

DOCUMENT RESUME

ED 266 681

FL 015 537

AUTHOR Kohonen, Viljo, Ed.; Pitkanen, Antti J., Ed.
TITLE Language Testing in School. AFinLA Yearbook 1985. No. 41.
INSTITUTION Finnish Association for Applied Linguistics, Helsinki.
REPORT NO ISBN-951-9388-28-1
PUB DATE 85
NOTE 252p.; Papers presented at a symposium of the Finnish Association for Applied Linguistics (Tampere, Finland, November 17-18, 1984) and a seminar on Language Testing in School (Heinola, Finland, November 19-21, 1984). For individual papers in English, see FL 015 538-548.
PUB TYPE Reports - Research/Technical (143) -- Collected Works - Conference Proceedings (021)
LANGUAGE English; Finnish
EDRS PRICE MF01/PC11 Plus Postage.
DESCRIPTORS Academic Ability; Communication Skills; *Communicative Competence (Languages); Educational Policy; Educational Research; Elementary Secondary Education; Foreign Countries; Language Research; Language Skills; *Language Tests; Policy Formation; Redundancy; *Second Language Instruction; Test Reliability; *Test Use; Test Validity
IDENTIFIERS Czechoslovakia; Finland; Norway; USSR; West Germany

ABSTRACT

Eighteen papers, 11 in English, and 6 in Finnish and 1 in German are presented on the general theme "Language--Knowledge and Skill," a theme that emphasizes the need to examine the communicative goal in foreign language education from the perspective of both theory and practice. The English-language articles on foreign language testing, following opening remarks at the Heinola seminar by Kalevi Pohjala include: "Criterion-referenced Testing in Foreign Language Teaching" (Sauli Takala); "Tests of Reduced Redundancy--Theory" (Christine Klein-Braley); "Tests of Reduced Redundancy--The C-Test, A Practical Example" (Raatz, Ulrich); "Testing Language Skills in the Norwegian Comprehensive School" (Lars Sigfred Evensen); "Effect of Mixing Ability Groups on Ability Levels Attained" (John H. A. L. de Jong); "Testing Communication Skills in Mixed-Ability Groups" (Viljo Kohonen); "On the Whys and Hows of Language Testing" (Torsten Lindblad); "Current Trends in FL Testing in Czechoslovakia and the Soviet Union" (Jan Prucha); "From Testing Research to Educational Policy: Introducing a New Comprehensive Test of Oral Proficiency" (Elana Shohamy, Thea Reves, Yael Bejarano); "Language Testing in German Schools--Theory and Practice" (Bernd Voss); and "The Teaching of Testing: Reliability, Validity, and All That Stuff" (Douglas K. Stevenson). (MSE)

BEST COPY AVAILABLE

SUOMEN SOVELTAVAN KIELITIETEEN YHDISTYKSEN
(AFinLA) JULKAISUJA

PUBLICATIONS DE L'ASSOCIATION FINLANDAISE
DE LINGUISTIQUE APPLIQUÉE (AFinLA)

No. 41

LANGUAGE TESTING IN SCHOOL
AFinLA YEARBOOK 1985

edited by

Viljo Kohonen and Antti J. Pitkänen

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

- Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

Tampere 1985

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. Tammola

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

ED266681

FL015537

Companion Volume:

Kahonen, Viljo, Hilika von Essen and Christine Klein-Braley (eds) 1985.
Practice and problems in language testing. B. AFInLA Series 40.
Tampere.

FOREWORD
CONTENTS

Lyle F. Bachman, An Examination of some language proficiency tests from a communicative viewpoint.....	5
Terry Culhane, Microcomputers and language testing.....	21
Peter Dickson, Evaluation of the productive skills in the second language the work of the assessment of performance unit (APU).....	27
Mary C. Kilborn, Developing classroom teachers' assessment skills.....	41
Christine Klein-Braley, Reflections on the content validity of holistic tests.....	53
Erkki P. Liski and Simo Puntanen, A percentile regression model for the number of errors in group conversation tests.....	63
Madelaine Lütjeharms, The development of reading comprehension in FL teaching - combining knowledge and skills.....	77
Kyle Perkins, Sheila R. Bratten, and Paul J. Angellis, Evidence for unique articulatory and auditory dimensions in ESL data.....	89
Jan Prucha, Foreign language needs: theory and empirical evidence in Czechoslovakia.....	105
Ulrich Rastz, Investigating dimensionality of language tests - a new solution to an old problem.....	123
Klaas Schreuder, Formative tests of speaking proficiency in the mother tongue for 14-year olds.....	137
Elana Shohamy, An alternative model for assessing language proficiency at the end of secondary school.....	149
Sauli Takala, Estimating students' vocabulary sizes in foreign language teaching.....	157
Robert Vanderplank, Evaluating the language laboratory in practice.....	167
Jan van Weeren, Language testing: making a decision on the basis of observed behaviour and 'tapped' knowledge.....	179
Pertti Widen, Cultural knowledge as part of language skills in cross-cultural communication.....	191
Bernd Voss, Classroom English for German schools.....	201

© The Finnish Association for Applied Linguistics (AFInLA)
and the authors

ISBN 951 - 9388 - 28 - 1
ISSN 0781 - 0318

FOREWORD

The present volume contains most of the papers presented on two occasions: the 1984 Symposium of the Finnish Association for Applied Linguistics (AFinLA) at the University of Tampere, November 17-18, and the seminar on "Language Testing in School", held at the Heinola Course Centre immediately thereafter, November 19-21. In connection with the AFinLA symposium, there was a parallel international language testing symposium of the Interuniversitäre Sprachtestgruppe (IUS) at the University of Tampere. The central theme of all the three occasions was "Language - Knowledge and Skill", a theme emphasizing the need to examine the current communicative goal orientation from the point of view of both linguistic competence and the ability - and the courage and willingness - to put that competence into pragmatically meaningful and socially acceptable use.

An important aim of the symposiums was to exchange information about current thinking and practices in various national contexts of language teaching; we have a great deal to learn from each other. An attempt was therefore made to obtain a wide international basis for sharing ideas. We wish to express our gratitude to the following institutions for their generous financial support: the America Center, the British Council, the Goethe-Institut, the Nordic Language and Information Centre, the Tampere University Foundation, and the Finnish Ministry of Education. Our thanks are also due to the University of Tampere for providing the facilities for the Tampere symposiums, and to the Heinola Course Centre for the warm hospitality at Heinola. Without all this support, the organization of the symposiums would not have been possible.

To disseminate information further to those who were not able to join us in Tampere and Heinola, most of the papers are published in two companion volumes in the AFinLA Series: the present one, and the IUS Symposium volume (Practice and problems in language testing 8); some of the papers will appear elsewhere in due course. We are grateful to Riitta Jaatinen and Jaana Rantanen for their help at the Tampere symposiums; to Ian Gurney for checking the English in some of the papers, and to Jouko Jousea and Riitta Könönen for their help with word processing.

Viljo Kohonen
Antti J. Pitkänen

CONTENTS

Kalevi Pohjala, Opening of the Heinola Seminar	5
Sauli Takala, Criterion-referenced testing in foreign language teaching	9
Christine Klein-Braley, Tests of reduced redundancy - theory	33
Ulrich Raatz, Tests of reduced redundancy - the C-test, a practical example	49
Lars Sigfred Evensen, Testing language skills in the Norwegian comprehensive school	63
John H.A.L. de Jong, Effect of mixing ability groups on ability levels attained	79
Viljo Kohonen, Testing communication skills in mixed- ability groups	91
Torsten Lindblad, On the whys and hows of language testing	109
Jan Prucha, Current trends in FL testing in Czechoslovakia and the Soviet Union	117
Elana Shohamy, Thea Reves and Yael Bejarano, From testing research to educational policy: introducing a new comprehensive test of oral proficiency	133
Bernd Voss, Language testing in German schools - theory and practice	145
Douglas K. Stevenson, The teaching of testing: reliability, validity, and all that stuff	165
Anneli Kauppinen, Äidinkielen taidon mittaaminen lukion päättövaiheessa	181
Heljä Kauhanen, Kommentti Anneli Kauppinen puheen- vuoroon äidinkielen taidon mittaamisesta lukion päättövaiheessa	189
Kyllikki Keravuori, Koululuokan diskurssin piirteitä	195
Matti Luukkainen, Wissen und Können im Erwerb der dritten und vierten Fremdsprache	209
Antti J. Pitkänen, Om språkliga drag och feltyper vid fri skriftlig framställning i främmande språk	227
Pauli Roinila, Tekstilajit ja kirjoitetun tekstin muotoutuminen	239
Vuokko Vuori-Barry, Kirjoittamisen opettamisesta lukiossa (A-englanti)	251

OPENING OF THE HEINOLA SEMINAR

Kalevi Pohjala (National Board of General Education)

Dear Friends,

I have the honour and pleasure to welcome you all to the Seminar on Language Testing in School at the Heinola Course Centre for the coming three days. This seminar is connected with the 8th International Language Testing Symposium and the AFinLA Symposium which were both held at the University of Tampere during the weekend. The theme of the Tampere conferences was Language - Knowledge and Skill.

As Finland had undertaken to arrange the 1984 Symposium the Finnish representatives of the IUS, Dr. Viljo Kohonen from the University of Tampere and Mrs. Hilikka von Essen from the Commercial College of Heinola, contacted me and asked whether the National Board of General Education was interested in arranging another seminar on language testing in Heinola. The National Board of General Education and its Course Centre in Heinola were naturally very glad about the offer and decided to take advantage of the presence in Finland of testing experts from so many countries as this would provide an excellent opportunity to compare current practices and trends in the field. In order to arrange this seminar on testing we formed a planning group which consisted of 5 persons: Viljo Kohonen, Hilikka von Essen, Lauri Ollila, Head of the Heinola Course Centre, Kalevi Peltola, the course assistant of the centre, and myself.

The planning group is grateful for financial support from the Finnish Ministry of Education, the British Council, the America Centre, the Goethe Institut, the Nordic Language and Information Centre and the University of Tampere. This assistance made it possible for us to invite 13 foreign guests. The seminar is thus an international event: there are participants from 11 countries.

The majority of the participants here, some 55, naturally come from our own country. The invitation to the seminar was sent to all teacher training departments and teaching practice schools of our universities, the Educational Research Centre of the University of Jyväskylä, the Matriculation Examination Board, all the 11 provincial governments, the Finnish Foreign Language Teachers' Association and about a dozen local educational authorities whose experimental schools have been involved in innovations. Through their representatives we will be able to disseminate the outcomes of the seminar widely in our country.

A conference like this is very important from the Finnish point of view at the moment. It is necessary to be aware of international trends in the field of foreign language teaching. Besides, we are undergoing a time of rapid changes in language teaching. Goals and contents of teaching are changing or have recently changed, new teaching methods are being developed and

the new technology is bringing new possibilities for language teaching. In view of such developments, it is necessary to evaluate the current practices in language testing.

In accordance with the new school law we are going to give up ability grouping in the teaching of foreign languages and mathematics in the comprehensive schools from the beginning of the next school year and we are going towards more heterogeneous teaching groups. To make this possible we revised our syllabuses for foreign languages in 1982. In these official documents emphasis has been shifted on to communicative language teaching. This means that evaluation will also be carried out according to one and the same syllabus for all learners. We have therefore had to experiment with new testing procedures in a number of schools. In those schools the emphasis has been transferred away from the comparison of students towards criterion-referenced assessment. Learning is in the first place evaluated with reference to the goals set for the teaching, and the students' results are compared with their earlier records. It is important to support the students in their learning task while it is less important to compare and classify them, though this is also needed to some extent. The task of the comprehensive school is above all to create such learning situations that every student can work as fully as possible according to his ability and conditions of life.

Post-comprehensive education, both in senior secondary schools and in vocational schools, is also undergoing major reforms affecting both the administrative framework and the contents of teaching. Both these reforms were similarly preceded by experimental work. As a result of the syllabus reform another big change can be expected in the late 1980's concerning foreign language tests in the matriculation examination, which students take at the age of about 19. The current test types include multiple-choice in reading and listening, a structural gap-filling test, and a short essay. Multiple-choice tests are difficult to set at suitable level of difficulty: some have been too easy, others too difficult. Besides, the theoretical basis of using the technique as a measure of communicative abilities can be questioned. A further problem is the backwash effect of testing on classroom practices. A lot of research and development work is being done in the field of testing both at the comprehensive schools and post comprehensive schools in Finland, and you will hear about these Finnish experiences during the seminar.

The seminar will focus on discussing the current state of the art of language testing within a communicative goal orientation in school. I hope the seminar will provide some answers to the following questions:

- current views of testing communicative competence
- current practices in matriculation exams - what can we learn from each other?
- criterion-referenced testing and graded objectives and tests for mixed-ability groups
- accuracy, fluency and comprehensibility in language use

- open-ended comprehension questions - how to ensure interrater reliability?

The international exchange of experiences and information is an important goal of the seminar. I believe the Heinola Course Centre will provide a pleasant environment for sharing information and learning from each other. To disseminate information further, our purpose is also to publish a report about the seminar in spring.

The programme is quite tight. We have as usual arranged some evening programme: the reception of the town of Heinola in honour of our foreign friends today, and tomorrow evening a visit to the new municipal theatre of Lahti. Those who want to see the first performance of the play *Lysistrata* are asked to sign their names on the list later today.

Monday will be a day of theoretical input, Tuesday will cover practical applications and solutions and Wednesday will be a time for summing up.

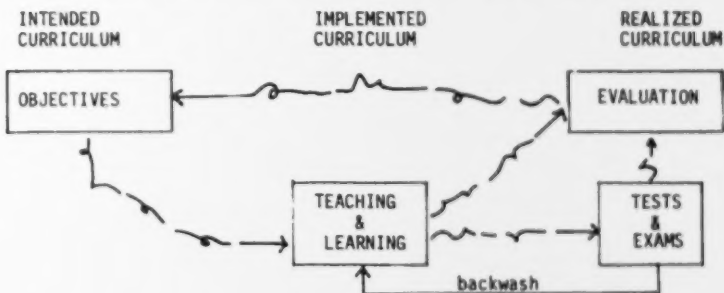
I hope you will enjoy your stay here and the conference will be a rewarding and stimulating experience to all of us. Once more, welcome to Heinola.

CRITERION-REFERENCED TESTING IN FOREIGN LANGUAGE TEACHING

Sauli Takala

1 Objectives - teaching - evaluations: the problems of match

One of the biggest problems in all educational activity is how to match objectives, teaching and evaluation. In other words, how the "intended curriculum" is transformed into the "implemented curriculum" and finally into the "realized curriculum" is no simple matter. Text-book writers and teachers interpret the intended curriculum and "implement" it to a certain extent; they cover some points thoroughly, some in a more cursory way, do not teach some points at all, and by contrast, teach some contents not mentioned in the intended curriculum. Students, in turn, learn ("realize") the implemented curriculum to a certain extent. The links between these three concepts are illustrated in the following figure.



The broken lines are intended to illustrate that there are no simple ways of getting from one stage in the process to another. There is a lot of interpretation and this means that there is a need to develop explicit and rigorous procedural rules to make such interpretation less ad hoc.

Tests and examinations are distinguished from evaluation. This is meant to convey the idea that evaluation is a broader (superordinate) term than testing. It is also meant to suggest that tests and examinations are not necessary for making evaluations. Evaluation means making judgements about the worth (merit, value) of something. In all walks of life we are constantly making such judgements. In education, evaluation usually means comparing what has been learned to what was intended to be taught.

Such judgements may be based on observation alone, but when properly made and used, tests can naturally give a valuable contribution to making correct evaluations. Tests are only one of the ways of getting data (information) for making evaluations.

2 Criterion- and Norm-Referenced Measurement

It was estimated that there were some 600 references on criterion-referenced measurement towards the end of the 1970's. Practically all of them were published during that decade. Yet, criterion-referenced measurement is not such a new idea.

E. L. Thorndike wrote about the difference between absolute and relative measurement some seventy years ago. Around 1950 Vahervuo in Finland carried out several studies on absolute and relative grading and on their theoretical basis. Still, it was in an article by Robert Glaser in 1963 that the term "criterion-referenced test" was introduced. The idea was favorably received but it did not lead to further work until in 1969 when Popham and Husek took up the concept and explicated further some of its implications.

Programmed learning and the behavioral objectives movement (e.g., Mager, 1962) were a major source in the emergence of criterion-referenced measurement. Carefully outlined teaching programs will not lead to a normal distribution of scores if the programs are, indeed, effective. There should be a high percentage of high scores and a decrease in variance. The latter is problematic for classical test theory, because most of its indices rely heavily on variance. Thus, it seemed necessary to conclude that variance-based estimates of test reliability are less appropriate in mastery-type instructional programs since they would unjustifiably label criterion-referenced tests as being of low reliability. New approaches were clearly needed (Popham and Husek, 1969).

Another major source, which is related to programmed learning and individualized learning programs, is the work done to discover learning hierarchies and curriculum (task) hierarchies (Gagne et al, 1962; Resnick, 1967). This work revealed that the testing of learning outcomes requires a thorough analysis of the subject matter as a preliminary step to item construction.

Criterion-referenced testing has been defined in a number of ways. According to Berk (1980a), at least fifty different definitions have been proposed since Glaser's first paper. Perhaps the most concise definition has been suggested by Popham (1978, p. 93): "A criterion-referenced test is used to ascertain an individual's status with

respect to a well-defined behavioral domain." This means that the interpretability of the test result is of primary concern. Whereas in norm-referenced measurement an individual's test score derives its meaning mainly from its relationship to the scores of other examinees (relative interpretation), the scores on a criterion-referenced test derive their meaning from the scores' relationship to a class or domain of tasks (absolute interpretation). Thus a domain score can be interpreted in terms of what an individual can do and what he cannot do and it also indicates what proportion of all possible tasks (items) of the whole item universe the individual could have solved if they were administered to him rather than only a sample of them. A domain score lends itself to absolute interpretations and can be used both for qualitative and quantitative descriptions (what is mastered and how much is mastered).

Several terms for this kind of testing have been proposed within the criterion-referenced movement. Ebel (1962) proposed a term "content-standard test" to describe a test which produces test scores which indicate what percentage of a systematic sample of defined tasks a person has solved correctly. Osborn (1968) used the term "universe-defined test" to refer to a test which produces an unbiased estimate of his score in an explicitly defined item content universe. Hively (1973) prefers the term "domain-referenced test" as a less ambitious term than universe-defined test. Carver (1974) has advocated the use of edumetric (rather than traditional psychometric) tests to measure within-individual growth (competence) instead of between-individual differences (ability, intelligence).

The term "objectives-based test" has sometimes been used as a near-synonym for criterion-referenced tests. If the items are simply derived from behavioral objectives without a strictly predetermined procedure, however, objective-based tests do not lend themselves to criterion-referenced interpretation.

The term "mastery test" has been derived mainly from the mastery learning system developed by Bloom (1968, 1971), largely on the basis of the model of school learning proposed by Carroll (1963). The main purpose of mastery tests is to help in the classification of students as masters or nonmasters of an objective in order to facilitate the management of an individualized teaching program.

If one were shown a test which only contained the instructions to students and the test items, it would be difficult to say whether the test is a criterion-referenced test or a norm-referenced test. In order to be able to make that decision it is necessary to know how the test was produced. It is in the work prior to the assembly of a test that most of the effort needs to be spent in producing a criterion-referenced test. Differences between two

forms of criterion-referenced testing (domain-referenced and mastery tests) and norm-referenced testing are summarized in Table 1. The first five stages in the development of tests refer to the planning stage and the rest to the technical aspects of tests and their uses.

3 Stages in test Construction

3.1 Specification of Content

It is in the specification of the content domain that the greatest challenge and also the greatest merit of criterion-referenced testing lies. In traditional norm-referenced tests the content limits are only partially specified. Short instructional and behavioral objectives are used as the basis for item generation. As Bormuth (1970) and Anderson (1972), among others, have shown, there is so much room left for interpretation that the items may reflect the characteristics of the test constructor more than those of the instructional program. Too much room is left for creativity, which according to Popham (1978, 1980), is not as desirable as strict adherence to the content limits. Several methods have been proposed for making domain specification more adequate. These will be discussed below in some detail, since this is a crucial part of all criterion-referenced measurement.

Item Transformations

Bormuth (1970) has suggested that linguistic analysis based on transformational grammar could be used to make explicit the methods by which items are derived from statements of instructional objectives. Bormuth advocates operationalism as a way of introducing rigor into item construction and sees syntactic operations as a promising way to do this. His method is illustrated below. It shows some item transformations that have been performed on a sentence "The older sister put out the fire." Using syntactic transformations several comprehension questions could be asked about the sentence.

Transformation Name	Question
Echo	The older sister put out the fire?
Tag	The older sister put out the fire, didn't she?
Yes-No	Did the older sister put out the fire?

Noun deletion	Who put out the fire? What did the older sister put out?
Noun modifier deletion	Which sister put out the fire?

It seems obvious that Bormuth's method is a useful tool for generating items testing the comprehension of written and spoken discourse. Anderson (1972) provides some other examples of ways of generating questions to test discourse comprehension. One weakness of these methods is, however, that the emphasis is on sentence level operations rather than discourse level units. Recent work on discourse analysis by Halliday and Hasan, van Dijk, Meyer and others will be of use in moving from sentence to discourse-level testing.

Mapping Sentence

Mapping sentences are used in facet analysis developed by Guttman (1969). Facet analysis can be used to describe the boundaries and structure of a domain of testing conditions. Facets are those dimensions or characteristics on which items in a given domain can differ. Facet analysis was used by the present writer in 1980 in an attempt to conceptualize the domain of written composition for the IEA International Study of Written Composition. The first attempt is illustrated below. (For a later version, see Takala, 1982.)

Millman (1978) also used facet analysis in his study of how the form and content of items are related to item difficulty.

Amplified Objectives

After finding out that item generation on the basis of traditional behavioral objectives was subject to too much interpretation and that using item forms was too demanding and led to "hyperspecificity", Popham (1980) worked with the so-called amplified objectives. As the name suggests, these are more detailed forms of behavioral objectives. They include 1) a brief statement of the objective, 2) a sample item, and 3) an amplified objective which specifies (a) the testing situation, (b) response alternative, and (c) criteria of correctness. The following example illustrates amplified objectives.

TABLE 1. Characteristics of Two Types of Criterion-Referenced Tests and of Norm-Referenced Tests (adapted from Millman, 1974, and Berk, 1980).

Stages of Development	Alternative Conceptualizations		
	Criterion-Referenced Testing		Norm-Referenced Testing
	Domain-Referenced	Mastery	
1. Specification of Content Domain	Maximum specification of content limits <u>Methods:</u> 1. Item transformations 2. Mapping sentences 3. Algorithms 4. Item forms 5. Amplified objectives 6. Test specifications	Content limits only partially specified <u>Methods:</u> Instructional and behavioral objectives	Content limits only partially specified <u>Methods:</u> Instructional and behavioral objectives
2. Item Construction	Generation rules	Traditional rules	Traditional rules
3. Specification of Item Domain	Infinite or finite item universe	Infinite ?	Infinite ?
4. Item Analysis	Purpose to detect flawed items <u>Methods:</u> 1. A priori judgment of item-objective congruence by subject matter experts 2. A posteriori computation of item statistics	Purpose to detect flawed items <u>Methods:</u> ?	Purpose to select items <u>Methods:</u> A posteriori computation of item statistics
5. Item Selection from Item Universe	Random	Nonrandom (?)	Nonrandom

TABLE 1 (cont.).

Stages of Development	Alternative Conceptualizations		
	Domain-Referenced	Criterion-Referenced Testing Mastery	Norm-Referenced Testing
6. Cut-off Score Selection	Optional	Required	Required (?)
7. Validity	Content Construct Decision	Content Criterion-related Construct Decision	Criterion-related
8. Reliability	1) Consistency of decisions (β_0, k) 2) Dependability ($\phi(\lambda)$) 3) Error of measurement or estimate around domain score using ϕ or other indices	Consistency of decisions (β_0, k)	Traditional procedures (based on correlation)
9. Score Interpretation	Performance in relation to domain (level of functioning) Performance in relation to required level of mastery	Performance in relation to required level of mastery	Performance in relation to other examinees
10. Item and Test Variance	Not required	Not required	Required

Mapping Sentence for the Domain of Writing Following Guttman's Facet Analysis Scheme

<u>A. Activity</u>	<u>B. Channel</u>	<u>C. Content/topic</u>	<u>D. Communication Partner</u>
1. Receive 2. Send	a/an	1. auditory message which deals with 2. visual	1. addressor 2. addressee
<u>E. Role relationship between addressor and addressee</u> has/is 1. a higher social status 2. an equal social status and which is 3. a lower social status 4. identical with addressor			
<u>F. Degree of publicity/formality</u> 1. private 2. semi-public 3. public			
<u>G. Input-output relationship (stimulus-response)</u> consisting of 1. repetition of input 2. modification of input and whose 3. internal input purpose is			
<u>H. Function</u> 1. to preserve the message (documentative) 2. to inform (referential) 3. to persuade (emotive) 4. to describe (descriptive) 5. 6.			

Different configurations of variables lead to different rhetorical modes (narrative, exposition, argumentation, etc.)

Examples:

A2 + B2 + C2 + D2 + E1 + F3 + G2 + H1 = a personal letter to a friend
A2 + B2 + C2 + D1 + E3 + F2 + G4 + H2 = letter of application

Objective: Given a sentence with a noun or verb omitted, the student will select from two alternatives the word which most specifically or concretely completes the sentence.

Sample Item

Directions: Mark an "X" through one of the words in parentheses which makes the sentence describe a clearer picture.

Example: The racer (tumbled, went) down the hill.

Amplified Objective

Testing Situation

1. The student will be given simple sentences with the noun or verb omitted and will be asked to mark an "X" through the one word of a given pair of alternative words which more specifically or concretely completes the sentence.

2. Each test will omit nouns and verbs in approximately equal numbers.

3. Vocabulary will be familiar to a third or fourth-grade pupil.

Response Alternatives

1. The student will be given pairs of nouns or pairs of verbs with distinctly varied degrees of descriptive power.

2. In pairs of verbs, one verb will either be a linking verb or an action verb descriptive of general action (e.g., is, goes), and one verb will be an action verb descriptive of the manner of movement involved (e.g., scrambled, skipped).

3. In pairs of nouns, one noun will be abstract or vague (e.g., man, thing), and one noun will be concrete or specific (e.g., carpenter, computer).

Criterion of Correctness

The correct answer will be an "X" marked through the more concrete, specific noun or through the more descriptive action verb in each given pair.

(Source: Millman, 1974)

While amplified objectives clearly define the measured domain and specify item generation in greater detail than simple behavioral objectives, Popham (1980) observes

that this attempt to "shoot for just the right balance between clarity and conciseness" failed. There was still too much room left for the personal interpretation of item writers.

Test Specifications

Experience with amplified objectives led Popham and his colleagues to believe that a so-called limited focus strategy was desirable. This means that the strategy is to focus measurement and to limit it to "a smaller number of assessed behaviors, but to conceptualize these behaviors so that they were large scale, important behaviors that subsumed lesser, en route behaviors" (Popham, 1980, p. 21).

The test specification consists of 1) a short general description, and 2) a sample item, which give the reader a general idea of what the test might contain. These are followed by 3) a detailed specification of the stimulus attributes and 4) response attributes including specification of the correct answer and, in the case of multiple choice items, of the reasons for various distractors. The test specification is illustrated below (Takala, 1984).

Domain Specification and Item Generation Rules for Vocabulary Size Assessment

Behavior

(1) When given a Finnish word in writing, the student can produce an acceptable English equivalent in writing (recall or active vocabulary). (2) When given an English word in writing, the student can produce an acceptable Finnish equivalent in writing (recognition or passive vocabulary).

Stimulus specification

The vocabulary presented in the core texts and extra (optional) texts in widely used English textbooks is listed. A stratified random sample is selected from the universe of such word lists. The words are presented without providing any context. Some of the words are used to measure both the passive and active knowledge of word meanings.

Response specification

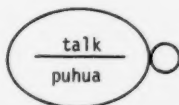
The student has to write the response in the space provided for that purpose.

Scoring

The responses are scored 0 - 1. A semantically acceptable and understandable response, which may contain spelling errors, is scored 1. In scoring active vocabulary, the decision is based on how the written English word would sound if read aloud. Thus, the student will get full marks if he/she has given the English equivalent of the Finnish word "talo" as "haus" instead of "house", since "haus" in Finnish orthography corresponds to the way "house" is pronounced in English.

Sample items

Instructions: "In this test you can show how well you know the English vocabulary included in your course work. Below are presented a number of Finnish words. Your task is to write the English equivalent on the line above the Finnish word. Write the word even if you may not be quite sure about the correct spelling, since spelling mistakes are a minor consideration in scoring."



"Write the Finnish equivalents of the following English words."



Popham (1980, 1981) feels that test specifications like the one shown in the above constitute a reasonable balance between clarity and conciseness so that busy people like teachers might not be put off by extreme specificity. Test specifications can also contain a supplement, which can give additional guidance in how to select stimuli, how to phrase questions, and so on.

3.2 Size of Domain

The proper size of the domain is, as so many issues in testing, ultimately dependent on the purpose of the test (measurement). A fairly large domain is appropriate if we are interested in more general forms of "terminal

behaviors" (i.e., we are doing "summative" evaluation and giving grades). A more limited domain definition is recommended when we are more interested in "en-route" behaviors and need information for deciding whether we need to review some matters with all students or give remedial help to some students ("formative" and "diagnostic" evaluation). Let us illustrate the issue of domain size with some concrete examples. Compare the following domain definitions.

1. Student can speak English
2. Student can ask for information in English
3. Student can ask about (a) time in English
(b) place
(c) cost
(d) another person's feelings
(e) another person's preferences
(f) another person's opinions
(g) another person's advice
etc., etc.
4. Student can ask for
(a) conformation of information (i.e., make yes/no questions)
(b) lacking information (make questions with HOW, WHAT, WHEN, etc.)
5. Student can make questions with WHAT
6. Student knows what WHAT means
7. Student knows how to spell/pronounce WHAT

Points 1 and 2 are probably too broad domain definitions to be of much use to language teachers. Levels 3 and 4 are probably of appropriate size for summative evaluation. Levels 5 through 7 might be considered for diagnostic purposes.

Following Baker (1974) and Popham (1975), it should be emphasized that we should not test trivial matters because that may lead to excessive testing. We should only test important matters and reserve as much time as possible for teaching and learning. Our tests should include only such items that contain those features and elements whose variation makes a difference in student response (Millman 1974). To take a simple example, it probably makes no difference whether we use "he" or "she" to test whether students can use the s-form in the simple present tense. If this is so, it is superfluous to have two items, one with "he" and one with "she".

Trying to concretize further the issue of domain size, it is probably not useful to treat as a domain something that can be taught and learned in one lesson. On the other hand, if something takes a whole term to teach, that probably is best divided into more than one domain.

dyna (1980; 1982) also provide a useful review of recent advances in the item-writing technology, including computer-based methods (cf. also Millman 1980). They note that the major positive result of the increased attention to the process of item writing is the heightened concern for the logical congruence between instruction and testing.

Once the rules for domain definition and for item generation have been worked out, it is necessary to consider specific items. Unlike in norm-referenced testing, it is necessary in criterion-referenced testing to know what the universe of items is that represents the defined domain content. This universe can be finite or infinite. As Millman (1973) points out, it is not necessary that the population of items actually exists. What is necessary, though, is that the domain is so well described that a high agreement can be reached about what items are and what are not members of the population.

Further, unlike in norm-referenced and mastery tests, it is necessary to draw a random sample from the universe of all possible items because only this procedure makes it possible to produce an estimate of the examinees' total domain scores. Random sampling of items is needed in order to make it possible to generalize into the whole domain tested. It is generally assumed that 10-20 items are needed to measure a given content domain.

5 Validity as an Issue in Criterion-Referenced Measurement

Criterion-referenced tests are more and more often used in monitoring individual progress through objectives-based instructional programs (formative testing), to diagnose learning problems (diagnostic testing), to evaluate educational and social programs (program evaluation), and to assess level of performance on certification and licensing examinations. The usefulness of such applications depends heavily on the validity of the procedures undertaken in such testing.

According to Hambleton (1980) validity considerations in criterion-referenced testing arise at three steps: 1) the selection of objectives (content domain), 2) the measurement of objectives (content domain) included in the criterion-referenced test, and 3) the uses of test scores.

Validity is a difficult topic in all measurement and criterion-referenced measurement is no exception. Terminology varies quite a lot so that different terms are used to designate the same characteristic and the same term is used to designate somewhat different thing. There are also some fundamental confusions that have persisted for a long time.

As Cronbach (1971), Messick (1975) and Linn (1979) have pointed out, a major conceptual confusion arises from the fact that content validity is focused on test forms rather than test scores, on instruments rather than measurements. In Linn's words "questions of validity are questions for the soundness of the interpretation of a measure ... Thus, it is the interpretation rather than the measure that is validated. Measurement results may have many interpretations which differ in their degree of validity and in the type of evidence required for the validation process" (Linn, 1979, p. 109). For this reason, Messick states that content coverage is an important consideration in test construction and interpretation but it does not itself provide validity. He would prefer the term "content relevance" or "content representativeness", since they do not really provide evidence for the validity of the interpretation of scores.

Popham (1978) uses the term "domain-selection validity" to refer to the question of how well the results obtained can be generalized to as many other domains as possible. It thus resembles "construct validity" to some extent, although the latter is a more theoretical concept. Since testing for many reasons ought to be limited to a minimum, it is important to measure such domains and use such techniques which permit maximum generalization across domains of content. Domain-selection validity can be assessed by asking experts to give judgements on the relevance of selected domains.

Popham (1978) proposes the term "descriptive validity" to indicate the representativeness of measured content. In traditional norm-referenced testing no quantitative indices are usually given to describe content representativeness (cf. Table 1). In criterion-referenced testing, judges can be used to assess to what extent items are congruent with the test specification. Hambleton (1980) provides some useful methods for doing this. In some areas, where it is possible to specify completely a pool of valid test items, the representativeness of items can be ensured by drawing a random sample from the item pool. This was the procedure adopted when the present author studied students' active and passive vocabulary of English in the Finnish comprehensive school in 1979.

Hambleton (1980) uses the term "decision validity" to refer to the decisions made on the basis of scores. Popham (1978) uses the term "functional validity" in much the same sense. Decision validity in criterion-referenced testing is often related to standard setting (minimum passing scores). Since that question is somewhat beyond the scope of this paper it will not be dealt with further in this context. A good review of decision-consistency is in Subkoviak (1980). Hambleton and Eignor (1978) and Walker (1978) review and assess standards and guidelines

for evaluating criterion-referenced tests and test manuals.

6 Reliability as an Issue in Criterion-Referenced Measurement

Traditional methods of estimating reliability in norm-referenced measurement are usually based on correlational analyses where variance is a key concept. Since there may be relatively little variation in the scores of criterion-referenced tests, correlation-based estimates may not be ideally suitable for the estimation of reliability.

As Berk (1980) has noted there are at least three major conceptualizations of criterion-referenced test reliability: 1) consistency of mastery-non-mastery decisions across repeated measures with one test form or parallel test forms, 2) consistency of squared deviations of individual scores from the cut-off scores across parallel or randomly parallel test forms, 3) consistency of individual scores across parallel or randomly parallel test forms.

Subkoviak (1980) gives a good survey of five methods of determining decision-consistency reliability. Usually only two statistics are used in this context: P_0 , which indicates the proportion of individuals consistently classified as masters and non-masters across parallel test forms, and κ , which estimates the proportion of individuals consistently classified beyond that expected by chance. Thus, P_0 estimates the overall consistency whereas κ estimates consistency due to testing alone. The choice of the index has to be based on whether one wants an estimate of overall consistency of decisions for whatever reason or of the contribution of the test alone. In most cases, it is probably advisable to report both estimates.

Brennan (1980) reviews the generalizability theory approach to reliability, which builds on the work by Cronbach and his associates (1972). Generalizability theory is based on the analysis of variance model and focuses on the estimation of various variance components in different types of test x items designs. Generalizability theory allows for the existence of many types and sources of error and it does not require strictly parallel tests for reliability estimation. Only randomly parallel tests are required.

As in the case of the decision-consistency approach, there are two indices of reliability (or dependability): $\phi(\lambda)$ provides an estimate of the dependability of mastery-non-mastery decisions based on the testing procedure λ (represents the cut-off score), and ϕ the "general purpose" index that is independent of the cut-off score and which

can be used to estimate individual domain scores (a major interest in the present writer's study of the size of students' active and passive vocabulary). $\hat{\sigma}(\lambda)$ is related to the reliability of criterion-referenced test scores and $\hat{\sigma}$ is associated with the reliability of domain score estimates. The former indicates how closely the scores for any examinee can be expected to agree, the latter the degree of agreement with chance agreement removed. Thus $\hat{\sigma}(\lambda)$ characterizes the dependability of decisions, or estimates, based on the testing procedure. Its magnitude depends, in part, on chance agreement. The index $\hat{\sigma}$ characterizes the contribution of the testing procedure to the dependability of decisions, over and above what can be expected on the basis of chance agreement (Brennan, 1980).

As in the case of the decision-consistency approach, it might be useful to give both estimates. Brennan (1980) also strongly recommends that variance components too should always be reported.

7 Discussion

Criterion-referenced measurement and norm-referenced measurement share a number of features. As in several other fields, for instance, in curriculum construction, new approaches usually mean only new emphases. At first there is a tendency to exaggerate differences. It is possible that this is inevitable when a new idea is introduced. Karl Popper has suggested that certain dogmatism may have an important part to play in the development of science, because giving up an idea too soon may mean that its merits and weaknesses are not given a sufficient chance of showing themselves. A scientist should not be too ready to adopt a new idea or to abandon an old one without persisting in some seemingly dogmatic stance for some time for the sake of argument. We should know how to play the believing and doubting games in a balanced way.

Criterion-referenced measurement shows some characteristics of this initial dogmatism. At first it was categorically stated that CRM does not need such concepts as item and score variance; that empirical item analyses are not needed; that norm data should not be gathered; and that content validity is the most important aspect of CRM. It was soon admitted, however, that these claims were overstated. Item variance usually occurs and serves a useful purpose in CRM testing as well as in norm-referenced testing. Similarly, it was conceded that norm data are not embarrassing for CRM. On the contrary, they add useful information and can help to interpret how "good" is "good enough". A posteriori empirical item analyses complement a priori judgemental (rational-logical) item analysis and help to detect flawed items. And, finally, content validity is not the all-important consideration in CRM.

While content representativeness is a necessary characteristic of CRM it does not guarantee the validity of interpretations based on CRT scores.

Criterion-referenced measurement has the special advantage that it provides an exact description of a person's performance level in an entire domain and not only on the presented items. Several requirements must be fulfilled before such an interpretation is possible. First, there has to be a detailed description of the measured domain. Second, there must be a detailed description of the instrument, which includes the specification of the stimulus and response parts and of the scoring system. Third, items must be generated that have a high item-objective congruence and which are also a representative random or stratified random sample from the item pool. If CRM is used for program evaluation there must also be a representative sample of students from the entire population. In the latter case it is advisable to use matrix sampling with several parallel test versions rotated in the class.

One of the greatest attractions of CRM for the present writer is its emphasis on the conceptualization of measured domains. This lends support to his personal claim, which goes back several years, that one of the greatest obstacles for the development of teaching is the lack of theoretically sound conceptualizations of the units and processes in learning a particular subject matter. He would, therefore, fully agree with the view recently put forward by Popham:

When created by instructionally astute developers, a criterion-referenced test can lay out so lucidly a set of teachable skills that the test itself becomes a potent force for instructional improvement. Instead of being an afterthought for use at the close of instruction, a properly conceptualized criterion-referenced test can stimulate measurement-driven instructional enhancement. Test developers can literally create test items so that they agree with one or more instructionally powerful explanatory constructs which teachers can then employ during their lessons ... This sort of focused instructional enterprise is not teaching-to-the-test in the negative sense that one teaches toward a particular set of test items. Rather, this approach constitutes teaching-to-the-skill, a highly effective and thoroughly defensible instructional strategy" (Popham, 1981, pp. 106-107).

Thus it might be that "the testing tail wagging the teaching dog" may not be such a problem or the embarrassment it is often taken to be if the tail is fully compatible with the dog. The present writer's personal experience with curriculum construction and evaluation, and with the in-service education of teachers in Finland suggests

that the most effective and fastest way to promote desirable changes in teaching is to make sure that testing and tests display the characteristics of desirable student performance. Tests are the most concrete ways of signaling to teachers and students what the desirable content and forms of learning are.

Focusing on testing may be more effective than focusing on curricula and teaching materials since testing has a more limited scope and it is, therefore, possible to produce very carefully constructed tests that are, in a sense, modules of teaching. Such tests can serve as examples for preparing units of teaching and for individual lessons. By concentrating on important aspects of the subject matter it is possible to produce such modules which can also serve as a stimulus for textbook writers. While individual units and modules do not constitute an entire syllabus, they are useful wholes as such and can serve as useful models. Practical experience shows that it is much more difficult to seek to conceptualize an entire curriculum with similar rigor and it is also a huge task to produce a textbook package with a similarly consistent approach. Thus testing may, indeed, be a sensible starting point and lead to improved curricula and textbooks. At the very least, the potential contribution of work done within testing and measurement to curriculum design and instruction should not be ignored.

If we continue to do serious work on testing, we can move from what some might describe as the modern "test cult" more and more towards "test culture". Test culture is characterized by several desirable features. First, there is an awareness of the importance of knowing why one is testing in the first place. Second, there is an awareness of the problems of how valid interpretations (conclusions) can be made on the basis of obtained, more or less reliable, scores. Third, there is an awareness of problems of the generalizability of the results to the whole content universe and to the whole student population. Fourth, and finally, there is an absence of dogmatism and taboos concerning test types. Test culture is mature when we are aware of every aspect of testing and evaluation being riddled with problems but we are only pleased to have been able to reach such a level of awareness.

References

- Anderson, C. 1972. "How to construct achievement tests to assess comprehension". Review of Educational Research 42: 145-170.
- Berk, R. A. 1980a. Introduction. In R. A. Berk (ed.) 1980. Criterion-referenced measurement: The state of the art. Baltimore and London: The Johns Hopkins University Press 3-9.
- 1980b. Item Analysis. In R. A. Berk (ed.) 1980. Criterion-referenced measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press 49-79.
- Bloom, B. S. 1968. Learning for mastery. Evaluation Comment, 1 (1).
- Bloom, B. S., J. T. Hastings and G. F. Madaus (ed.). 1971. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill.
- Bormuth, R. 1970. On the theory of achievement test items. Chicago: University of Chicago Press.
- Brennan, R. L. 1980. Applications of generalizability theory. In R. A. Berk (ed.). 1980. Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press 186-202.
- Carroll, J. B. 1963. "A model of school learning". Teachers College Record 64: 723-733.
- 1968. The psychology of second language testing. In a Davies (ed.) 1968. Language testing symposium: A psycholinguistic approach. London: Oxford University Press 46-68.
- 1976. Psychometric tests as cognitive tasks: A new "structure of intellect". In L. B. Resnick (ed.). 1976. The Nature of Intelligence. New York: Lawrence Erlbaum 27-56.
- Carver, R. P. 1974. "Two dimensions of tests: Psychometric and edumetric". American Psychologist 29, 512-518.
- Cronbach, L. J. 1957. "The two disciplines of scientific psychology". American Psychologist 12: 671-684.
- 1971. Test validation. In R. L. Thorndike (ed.). 1971. Educational Measurement. Washington, D.C.: American Council of Education 443-507.

Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam. 1972. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Ebel, R. L. 1962. "Content standard test scores". Educational and Psychological Measurement 22: 15-25.

- 1972. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall.

Gagne, R. M., J. R. Mayor, H. L. Garstens, and N. E. Paradise. 1962. Factors in acquiring knowledge of a mathematical task. Psychological Monographs Vol. 76, No. 526.

Glaser, R. 1963. "Instructional technology and the measurement of learning outcomes: Some questions". American Psychologist 18: 519-521.

Glaser, R., and A. Nitko. 1971. Measurement in learning and instruction. In R. L. Thorndike (ed.). 1971. Educational measurement. Washington, D. C.: American Council on Education. 652-670.

Guttman, L. 1969. Integration of test design and analysis. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service.

Hambleton, R. K. 1980. Test score validity and standard-setting methods. In R. A. Berk (ed.). 1980. Criterion-referenced measurement: The state of the art. Baltimore and London: The Johns Hopkins University Press. 80-123.

Hambleton, R. K., and D. R. Eignor. 1978. "Guidelines for evaluating criterion-referenced tests and test manuals". Journal of Educational Measurement 15: 321-327.

Hambleton, R. K., H. Swaminathan, J. Algina, and D. B. Coulson. 1978. "Criterion-referenced testing and measurement: A review of technical issues and developments". Review of Educational Research 48: 1, 1-47.

Hively, E., G. Maxwell, G. Rabehl, D. Sensen, and S. Lundin. 1973. Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project. CSE Monograph Series in Evaluation, No 1. Los Angeles: Center for the Study of Evaluation, University of California.

Linn, R. L. 1979. Issues of validity in measurement for competency-based programs. In M. A. Buda and J. R. Sanders (eds.). 1979. Practices and Problems in Competency-Based Measurement. Washington, D. C.: National Council on Measurement in Education. 108-123.

- Lord, F. M. 1976. Test Theory and the Public Interest. Proceedings of the 1976 ETS Invitational Conference. Princeton, N.J.: Educational Testing Service.
- Mager, R. F. 1962. Preparing instructional objectives. Palo Alto: Fearon Publishers.
- Messick, S. A. 1975. "The standard problem: Meaning and values in measurement and evaluation". American Psychologist 30: 955-966.
- Millman, J. 1973. "Passing scores and test lengths for domain-referenced tests". Review of Educational Research 43: 205-216.
- 1974. Criterion-referenced measurement. In W. J. Popham (ed.) 1974. Evaluation in education: Current applications. Berkeley: McCutchan. 310-397.
- 1978. Determinants of item difficulty: A preliminary investigation. Center for the Study of Evaluation. CSE Report No. 114.
- 1980. Computer-based item generation. In R. A. Berk (ed.). 1980. Criterion-referenced measurement: The state of the art. Baltimore and London: The Johns Hopkins University Press. 32-43.
- Osborn, H. G. 1968. "Item sampling for achievement testing". Educational and Psychological Measurement 28: 95-104.
- Popham, W. J. 1978. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice Hall.
- 1980. Domain specification strategies. In R. A. Berk (ed.). 1980. Criterion-referenced measurement: The state of the art. Baltimore and London: The Johns Hopkins University Press. 15-31.
- 1981. Measurement essentials for the essentials of education. In L. Y. Mercier (ed.). 1981. The Essentials Approach: Rethinking the Curriculum for the 80's. U.S. Department of Education. 97-115.
- Popham, W. J., and T. R. Husek. 1969. "Implications of Criterion-Referenced Measurement". Journal of Educational Measurement 6: 1-9.
- Roid, G. H., and T. M. Haladyna. 1980. "The emergence of an item-writing technology". Review of Educational Research, 50: 293-314.

Roid, G. H. & Haladyna, T. M. 1982. A technology for test-item writing. New York: Academic Press.

Subkoviak, M. J. 1980. Decision-consistency approaches. In R. A. Berk (ed.). 1980. Criterion-referenced measurement: The state of the art. Baltimore and London: The Johns Hopkins University Press. 129-185.

Takala, S. 1982. "On the Origins, Communicative Parameters and Processes of Writing". Evaluation in Education 5: 209-230.

Takala, S. 1984. Evaluation of students' knowledge of English vocabulary in the Finnish comprehensive school (Tech. Rep. No. 350). University of Jyväskylä. Institute for Educational Research.

Walker, C. B. 1978. Standards for Evaluating Criterion-Referenced Tests. Center for the Study of Evaluation, CSE Report No. 103.

CONTENTS	LANGUAGE KNOWLEDGE			LANGUAGE SKILLS		COMMUNICATION SKILLS (LANGUAGE USE SKILLS)			
	Levels of language/ discrete tests			Sub-skills/ integrative tests		Pragmatic tests			
	PHONOLOGY	MORPHOLOGY & SYNTAX	SEMANTICS	Listening comprehension	Speaking	Reading comprehension	Writing	Non-verbal	Verbal
MECHANICAL SKILLS (discriminate, repeat, recognize, recall)	Which is different? Which are identical? Imitate! Repeat!	Recall conjugation, paradigms Do in accordance with the model! Recite a rule	Choose/ give -I1 equivalent -I2 equivalent	Write down phone numbers Spell names	Imitate! Repeat! Respond to..	Which is different? Which are the same?	Copy! Which are the same?	Conventional gestures and facial expressions in conversational interactivity Interpret! Use gestures ..!	Conventional verbal expressions customary in verbal interaction Interpret! Act (speak/write) in a manner expected in the context!
	Minimal pairs	Multiple choice Matching Reorganization	Multiple choice Matching					Eg. smiling (What were going to say?) or a gesture with the thumb (In that direction)	Dictation Proof-reading Linguistic rituals (Filters etc)
KNOWLEDGE UNDERSTANDING (recognize, recall, classify, state...)	Is x voiced/ voiceless? Is x short or long? Where is main stress? What intonation?	What tense? Sing/plural? What case? What does x refer to? Which structure fits the context?	What word class? What is the meaning of x (using word formation)?	What was said? What facts were stated? What opinions were expressed?	State/present/ inform/express..	What was written? What facts/ opinions were expressed?	Write, state, present, express	Respond/act in a way customary in context	Respond and act (speak/ write) in a way customary in the context
	Multiple choice Matching			True-false Multiple ch. Matching Classification Question-answer Error spotting	Reading aloud Guided speaking	True-false Multiple choice Matching Classification Question-answer Error spotting	Fill-in Transform. Guided writing	Eg. movement of hand with the palm upwards (Ladies first! Would you like to say something as this point?)	Taking notes Interpreting Chatting Reporting Close
APPLICATION (perform, execute, demonstrate, produce, invent, create..)	Read aloud	What structure would you use in this context?	What is the corresponding adj., adverb, verb etc of this word?	What does the speaker really wish to communicate?	Produce, create, express	What does the writer really want to say?	Produce, create..	Respond and act creatively	Respond and act creatively
		Fill-in Transformation Multiple choice	Quest-answer Fill-in Multiple ch.	Selection of gist Making inferences Correcting errors Criticizing Commenting	Role playing Oral presentation	Underlining main points Making inferences Correcting errors Criticizing text	Reproduction	Eg. smiling in this context. X says to Y: "How about some coffee?" Y smiles and transforms implied request into an offer	Interview Conversation Summary Narrating Free writing Fluent translation

TESTS OF REDUCED REDUNDANCY - THEORY

Christine Klein-Braley (University of Duisburg)

0 Introduction

Scientific language testing is generally thought to have started in 1961 when Robert Lado published his book Language Testing, and John B Carroll his paper Fundamental considerations in testing for English language proficiency of foreign students. The tests which were developed as a result of these impulses consisted of multi-item subtests for the various aspects of language: grammar, vocabulary, phonology, reading and listening comprehension and so on. They conformed to all the demands of psychometrics. Since the items used were generally cast in multiple-choice form they were objective. Test development procedures using item analysis ensured high reliability. Correlations with other criteria such as teacher judgements showed their empirical validity.

Nevertheless in the early seventies linguists were becoming increasingly dissatisfied with this type of test. The multiple-choice item, they said, tested recognition rather than production, and there was no guarantee that a student who could identify the correct solution would necessarily be able to use the form in productive speech. Isolation of problems in individual items provided students with a clue as to what was being tested at any given moment. There seemed to be little theoretical justification for the implicit weightings accorded to the various subskills tested: a test which assigns twice as many points for vocabulary as for grammar makes the assumption that vocabulary is exactly twice as important in language, but there was no way of testing whether this assumption was true. In addition, the intercorrelations between subtests often turned out to be so high that - at least statistically - there was no real difference between the areas they were allegedly testing. The major criticism, however, was that such tests bore very little relationship to real-life language use: as Spolsky pointed out in 1973 they fail to take into account two vital truths about language, the fact that language is redundant, and the fact that it is creative.

1 Information theory and language

One of the solutions proposed to this problem was the construction of tests of reduced redundancy. The theoretical justification for such tests was put forward by Spolsky in 1973, and was based on Shannon and Weaver's communication theory (1949), nowadays generally referred to as information theory.

This theory was originally developed under the auspices of the Bell Telephone Company, and is concerned with the efficiency with which a message is transmitted from a source to a receiver. The content of the message is the information, although it is important to realise that the concept of information, as it is used here, has nothing to do with the meaningfulness of the message. Information has to be looked on as something more abstract, like weight for instance. The information is encoded into symbols which may be numbers, letters, or, as in morse code, bleeps. At first it would seem that the ideal code for any message is one in which each symbol transfers the maximum amount of information. Numbers are like this. If a telephone number is to be transmitted, then every single digit is equally important: none can be missed out and none can be changed. No digit in the string can be reconstructed from knowledge of the ones preceding or following it. So numbers transmit the maximum amount of information. But because of this great care is needed, otherwise one can easily find that one has dialled Buenos Aires (from Germany 00541 ...) when one actually intended to dial London (00441 ...). Similarly, the letters B, P and V sound very alike in English and can easily be confused by the listener.

Any disturbance in communication is technically known as noise. Noise can occur at any point in the communication process: the typist can hit the wrong key, there can be atmospheric or electrical disturbances in the telephone line, the receiver can misread or mishear what arrives at his end. Noise is unpredictable and random and there is no way of telling whether a given message will be affected by it. But a message which arrives at the receiver so damaged that no sensible interpretation is possible is a waste of time and resources, and so noise must be countered if this is feasible.

For this reason communications engineers have to build safeguards into the system, and since it is very expensive to remove all technical sources of noise, they prefer to use more symbols than are strictly necessary and to ensure that the coding is performed in such a way that errors caused by noise can be minimised. For instance, both bank account numbers and the international standard book numbers include check digits. When the numbers are entered into the computer some sort of arithmetical manipulation is performed, and if the result of the computation is not confirmed by the check digit, then the input is rejected as incorrect. If one is dictating a telegram, the post office asks one to use disambiguating names: "B for Bertha, P for Peter, V for Victory" etc. In this way, although it takes longer to dictate the telegram, one can be sure that the message has been taken down correctly.

The additions inserted into the message as safeguards are known as redundancy, and in terms of information theory redundancy is desirable. Shannon and Weaver's motivation for developing communication theory was to work out the mathematics of the process and thus to enable communications engineers to find the optimal balance between noise and redundancy in the transmission of messages - as much redundancy as necessary but as little as possible.

If we look at natural language in terms of information theory we find that it is a highly redundant system. Shannon estimated that printed English (in terms of letters) is about 75 % redundant. This means that the same amount of information could be conveyed by using only 25 % as many letters and spaces as at present. This is the technique adopted by the Speedwriting form of shorthand: most of the vowels are omitted, and some of the consonants, and yet the message can still be read:

gt a gd job & mo pay

say the advertisements for Speedwriting courses in the London tube.

As Spolsky points out, when one considers all the interferences that occur when natural language is used for communication, it is clear that only a redundant system would work. When we are confronted with bad handwriting, imperfect photocopies, noisy telephone lines, station loud speakers, it is redundancy which enables us to make sense of what we hear or read, even though some elements of the message are missing.

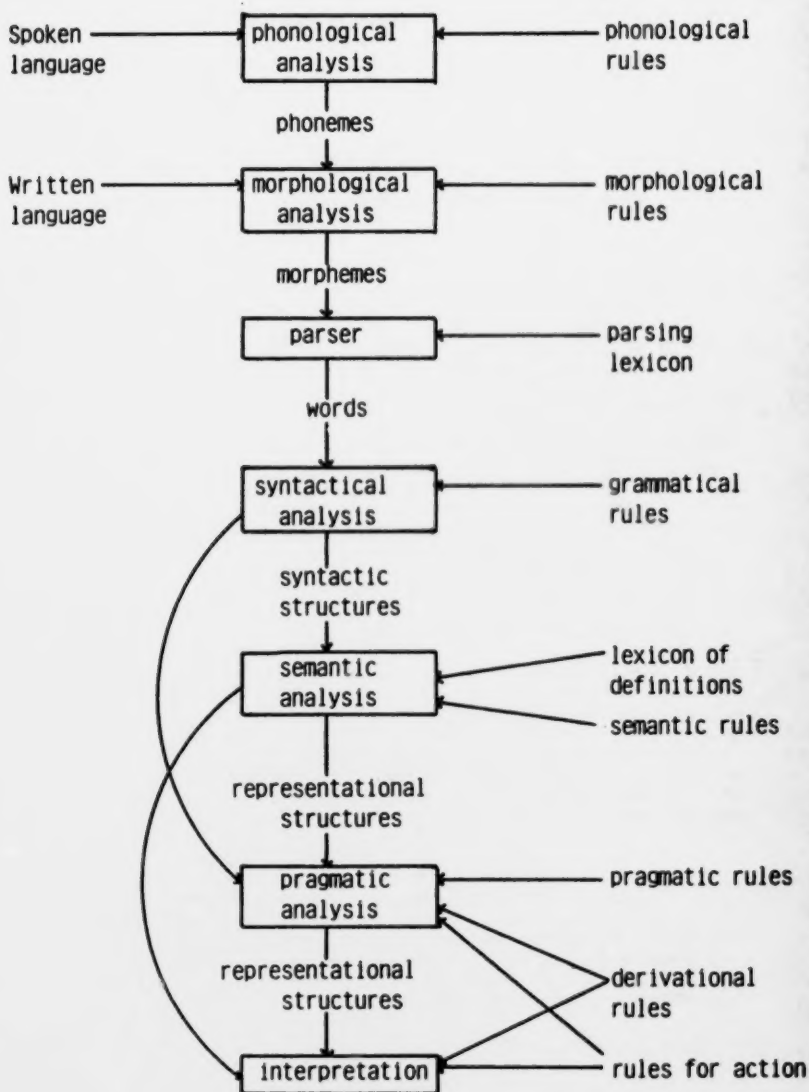
Moreover redundancy is present on a variety of different levels. Language is a very highly structured system consisting of inter-related subsystems, and we work with all these simultaneously. Just how much goes on when we are processing language becomes obvious if we look at the instructions which have to be given to that electronic idiot, the computer, if it is to understand and act on a simple English sentence (Diagram 1, taken from Winograd 1984). But we can perform this analysis without even thinking about it - because we know the rules of the language.

2 Redundancy and language testing

While native speakers of a language can use its redundancy to supply missing elements and thus restore the original message, non-natives need the full normal redundancy, and sometimes even more. Spolsky points out that when we talk to a foreigner we speak more slowly, more clearly, with added gesture and frequent repetition. He had been able to show (Spolsky et al. 1968) that in dictation tasks a clear distinction between natives and non-natives became evident as soon as noise was added. He explains that this is a result of:

the non-native's inability to function with reduced redundancy, evidence that he cannot supply from his knowledge of the language the experience on which to base his guesses as to what is missing. In other words, the key thing missing is the richness of knowledge of probabilities - on all levels, phonological, grammatical, lexical, and semantic - in the language. (1973:170).

-36-
Diagram 1: Computer analysis of real language input



In his view two implications for language testing follow from this:

1. It is problematical to try and decide that someone knows language on the basis of his performance on discrete-point items since it is impossible to demonstrate that any one item is essential for successful communication.
2. Knowing a language certainly involves the ability to understand a distorted message, to make valid guesses about a certain percentage of omitted elements.

In order to make valid guesses about missing elements a language user has to know the rules of the language. In Chomsky's terms he has to activate his competence (Chomsky 1965) in order to produce an acceptable performance. Spolsky therefore come to the conclusion that tests of reduced redundancy are techniques for making a learner "perform his competence".

Spolsky also points out that linguistic performance may be either active or passive: both the speaker and the listener are performing. Linguistic competence underlies both kinds of performance. The implications of this for language testing are important because it means that a test of the receptive skills can reveal competence just as well as a test of the productive skills.

A test of reduced redundancy, therefore, is one which introduces noise into the channel and asks the learner to perform in spite of this noise. The noise can be acoustic, but it can equally well be induced by deleting portions of a text.

3 Operationalisation of the theory

In order to convert a theory into a test a number of practical decisions must be taken, in other words the theory must be operationalized. A number of different operationalizations have been proposed in order to make use of reduced redundancy in language testing. They differ in a variety of aspects (see Table 1):

1. the definition of "text";
2. the definition of "noise";
3. the choice of channel;
4. the type of behaviour expected from the learner.

Table 1
Operationalization of reduced redundancy

TEST	text	noise	channel	examinee behaviour
Noise Test	single sentences	white noise at varying levels of intensity	acoustic	writing down what was heard
Partial dictation	2 longer texts	deletion of phrases selected by test constructor	acoustic/visual	writing down what was heard in blanks (examinee provided with mutilated text)
Cloze test	1 longer text	deletion of words (every nth)	visual	filling in blanks in written text

However, in all cases they share the assumption that the performance shown on the text can be linked to the learner's competence, if not absolutely then at least in terms of rank ordering of the examinees, i.e. the examinee with the higher score has a more highly developed competence than the examinee with a lower score.

In the discussion which follows I should like to examine briefly three different operationalizations of the theory of reduced redundancy and discuss the implications of the results obtained with these tests for the original theory.

3.1. The Noise Test

The first studies with the Noise Test were reported in 1968 by Spolsky et al. The basic idea was to dictate short sentences to the examinees under varying conditions of noise. The examinees were to write down what they heard, or thought they heard.

A variety of preliminary experiments were conducted, and the final 1968 version of the test consisted of 50 sentences, controlled for syntactical complexity in terms of the number of transformations required, and controlled for vocabulary in terms of word frequency. White noise - which sounds like a constant hissing - was added acoustically at signal to noise ratios of 3 decibels (the highest noise level), 4, 7, 10 and 50 decibels. To avoid learning effects the highest noise levels were administered first. 5 points were given

for a correct response, 4 for a response with one phonemic error, 1 for a response with at least one word correct and 0 otherwise.

In all the early trials the test had proved to discriminate very well between native and non-native speakers (with the non-natives achieving lower scores). The final version was administered to 61 foreign students who entered Indiana University in 1966. The results of the Noise Test were correlated with the results of other English tests administered by the EFL department and a comparison of the scores with the levels of proficiency assigned by the EFL department (based on the test scores of their own tests and a short interview) showed that the Noise Test successfully identified the high and low scorers. There were some problems however in the middle range.

Table 2

Correlations between Noise Test and EFL tests

	IU Aural Comprehension	Objective	Essay
Noise test (B4)	.66	.66	.51
IU Aural Comp		.61	.62
Essay			.65

In 1975 Gradman and Spolsky reported that a multiple-choice version of the Noise Test had been produced and that the correlation between the original and this version was .89 for one group of non-natives and .86 for a second. The test had also been administered to 26 Saudi-Arabian students who took the TOEFL at the same time. Here the correlation was .75 with the total TOEFL, and with the exception of English Structure and Writing (.44 and .33 respectively); all subtest correlations were above .60. In addition a correlation of .69 emerged between the Noise Test and the Ilyin Oral Interview. Similar correlations emerged for a second group of 25 Saudi students. For a group of 71 foreign students the correlation between the Noise Test and the Indiana University Placement Test was .56.

However there was a fly in the ointment. In the process of construction validation, i.e. in the attempt to relate test results and theory, the scores on the last 10 sentences, which were accompanied by a very high signal/noise ratio (50 db) and which therefore seemed to have no noise overlying them, were correlated with the other 40 test items. The correlation proved to be .93, "which may lead one to believe that as an overall measure of language proficiency, the noise test might just as well be given as a dictation test without the added distorting noise" (Gradman/Spolsky 1975: 65). This suspicion hardened when other noise levels were examined: only at the 7 db level and below did the distorting noise prove more advantageous to the native speaker.

Examination of the test sentences revealed that those sentences on which subjects did least well were the ones for which comprehension seemed to be most dependent on context. The researchers therefore concluded that in these cases reduced redundancy was indeed operating, but it was redundancy of context which was missing and which most hampered non-native performance. A contextualized version of the Noise Test was therefore developed in two parallel forms with 20 items each. The test material consisted of short dialogues and mini-lectures of the type which might be encountered in the classroom or in everyday life. This time pink noise was added.

The tests still discriminated well between natives and non-natives, and the natives had mean scores of 99 %. However, when the data were analysed in detail it was impossible to show that the noise had had any effect on the scores. The mean difference between scores with and without noise was only 3.9 points for Form A and a mere 1.6 points for Form B. The rank correlations for the 10 sentences with noise and the 10 without were .97 on Form A and .91 on Form B for the non-natives. Reluctantly, but honestly, the researchers came to the conclusion that: "the use of background noise seems to have had little effect on the measurement of overall proficiency for our subjects" (Gaies et al. 1977: 56).

We can therefore conclude that this attempt to operationalize reduced redundancy is not a viable one in its present form (1).

3.2 Partial dictation

Johansson's operationalisation of the concept of reduced redundancy was the partial dictation test. As he points out himself (1973b:2) it is closely related to two other measures thought to be tests of general language proficiency: dictation and cloze tests (2). Johansson reviews the work of Oller (1971) and Valette (1967), both of whom come to the conclusion that dictation is a satisfactory test of overall language proficiency, but he seems to feel that normal dictations are too easy for advanced students. Partial dictations, which are based on extremely difficult texts, are therefore a suitable technique.

Such tests have a number of advantages, claims Johansson. It is possible to concentrate on selected portions of the text. The passage is played from a tape recorder, which means that a variety of different voices and speech situations are possible. Students listen to the recordings only once, so that partial dictation is more economical.

Johansson describes two different techniques for developing his dictations. In the first case authentic material is recorded and then transcribed, in the second a suitable text is selected and then read on to tape by a native speaker. In both cases the test constructor decides what is to be omitted and ensures that the necessary pauses are patched into the tape immediately after the item deleted in the student text. The decision about what to delete is left to the test constructor's imagination. Technically it is advisable to delete segments of language which come before a natural pause in the text. Johansson's deletions vary between 2 and 10 words. Administration in

a language laboratory is advisable.

The tests are scored by assigning one point for every correct word restored. Spelling errors which do not interfere with understanding are not counted.

Johansson reports a correlation of .88 (N=80) between his two texts (parallel test reliability) and high correlations between these test scores and other tests (Table 3):

Table 3
Correlations between partial dictations and other tests

English vocabulary	English vocabulary and grammar	English pronunciation	Swedish vocabulary	English speaking test
.80 (N=20)	.86 (N=20)	.78 (N=20)	.65 (N=20)	.62 (N=36)

However there are two puzzling things about these statistics. The first is why the full correlation matrix is not presented - this is normal practice since it enables the reader to examine patternings in the data for him/herself. Mean scores and standard deviations are also missing. The second curiosity is the inclusion of only 20, allegedly randomly selected, students in the correlations when 80 students took the test.

So far as the statistics go, this approach seems to work satisfactorily for these two texts and this sample. However there are problems. Spolsky et al. were working with a specific test, and at no point in their discussions do they imply that their results are to be generalized to other test versions. Johansson makes precisely this claim - "due to the high correlations with other measures of foreign language proficiency, it might be used as a global estimate of a subject's proficiency in a foreign language" (1973b:42). This is a dangerous recommendation to make on the basis of only one test run and an analysis of only one quarter of the test subjects, particularly since there is no attempt made at any point to provide some sort of theoretical framework for the deletion techniques used.

Nor is any attempt made to link up the results obtained to a general theory of reduced redundancy. One possible approach, for instance, would have been to use the two other tests explicitly mentioned in the paper, cloze and dictation, as validation criteria. Or something along the lines of the multitrait-multimethod matrix would have been possible, using the integrative measures on the one hand and the discrete-point tests already available on the other.

In other words, there is no reason to assume at this point that partial dictation is a valid test of reduced redundancy, nor that

it measures general language proficiency: We just don't know!

3.3 Cloze tests

The most important and well-known operationalization of reduced redundancy is, of course, the cloze test. The test - and the name - were first used by Taylor (1953). It consists, as everyone probably knows, of a piece of prose in which every n th word been deleted (where n is a number between 5 and 10). The examinee restores the text by filling in the missing words. Scoring is either exact, so that only the words which the original author used are counted as correct, or it is acceptable, and in this technique any word which fits into the text semantically and syntactically scores a point.

The difference between the cloze test and other gapfilling procedures is the explicit link-up between deletions and information theory. The systematic deletions are intended to function in the same way as random noise in the communications system - Taylor's pilot studies had allegedly shown that systematic deletion was a good approximation to random deletion using random number tables.

Taylor's original aim was to develop a new readability technique, and in his experiments it was the texts, not the subjects, which were being tested. However he did, in fact, find significant differences between individuals and therefore suggested that cloze tests could be used to measure reading proficiency in the native language (L1).

In 1959 Carrol, Carton and Wilds experimented with cloze procedure for foreign language testing, but their results were negative. In 1968 Darnell obtained positive results with cloze tests for non-native speakers of English, but since his scoring procedures involved first testing a group of natives and then assigning points for the non-native responses on the basis of the frequency with which the particular response had been offered by the natives (clozentropy) this was not a technique which could be advocated for general use. However in 1971 Oller and Conrad performed an investigation using exact scoring, and produced results which showed that natives scored higher than non-natives and that the mean scores of the five non-native groups faithfully reflected their teaching classifications (from beginners to graduates in TESL). Furthermore the cloze scores showed a multiple correlation with the UCLA ELSPE test of .88. Oller and Conrad concluded: "This pilot study lends further support to the applicability of the cloze technique to the problems of ESL proficiency testing" (1971:192).

Oller continued to work on cloze tests. In 1972 he examined questions of difficulty levels and scoring techniques, and concluded that for all three levels of difficulty acceptable scoring was better: the correlations with the validation criteria were higher and the tests were more reliable. In 1973 he turned his attention to the question of what cloze tests measure:

On the basis of incomplete information, the subject is required to project a word to fill in a blank and thereby complete a sentence. In doing so, both the so called "productive" (or "active")

and "receptive" (or "passive") skills of a language must be utilized. The information provided in the cloze test allows the student by analysis to synthesize a greater whole. At the same time, the synthesis or projection may become part of the next analysis required to produce a subsequent synthesis.

It is interesting to note that the process ... involves more than "passive" reading. By sampling the information that is present the subject formulates hypotheses, or expectations, about information that is to follow. By sampling subsequent sequences, he either confirms or disconfirms these expectations. If the expectations are disconfirmed they must be revised and new hypotheses must be formed ... I would like to suggest that they are ... manifestations of the ... underlying competence. (1973:114).

He coined the term pragmatic expectancy grammar, and linked it to Miller's concept of "chunking" information to explain the subjects' behaviour (Miller 1970:20 ff.). Miller had pointed out that more material could be retained in short-term memory if it could be "chunked" or structured, and had also shown that grammatical structuring was a very efficient method of "chunking". All this, of course, fits in very well with Spolsky's original theory for tests of reduced redundancy.

Empirically, cloze tests repeatedly showed excellent performances. As measured by KR 20 they were highly reliable, and they showed high correlations with the validation criteria (usually discrete-point tests). Furthermore the highest correlations were generally found between cloze tests and other holistic approaches to proficiency testing: listening comprehension tests and dictations, for instance. (For overview, see Oller 1979). It came to be generally accepted that cloze tests were reliable and valid tests of general language proficiency. They were recommended for use by teachers: they were easy to construct, to score, particularly by the exact method, and their reliability and validity were known to be high.

It is important to realise, however, that the claims for high reliability and validity are based on theoretical assumptions:

1. The learner has reached a certain level of competence in the foreign language, i.e. has acquired a structured international grammar which is activated in all use of the language.
2. Cloze tests force the learner to use his internal grammar to restore missing elements.
3. The score obtained on a cloze test is a numerical estimate of the efficiency or quality of the learner's grammar.

This means that the text in a cloze test has a specific function: it is a sample of the language. The blanks are then a sample of the text, and these in their turn are intended to sample the learner's grammar. Only by making these theoretical assumptions can we legitimately state, as Oller, for instance, does, that the specific text involved is not important; any text will do (1979:364). This is also the reason that

deletion rate is not important: what is actually intended is random deletion, and random sampling is the best way of ensuring a representative sample of whatever we wish to measure.

The most obvious way of testing these assumptions is to examine the equivalence of cloze tests. If the theory is correct then the following statements should be true:

1. Random deletion of words and nth deletion of words produce equivalent samples of the text.
2. Tests constructed from the same texts but using different deletion rates or starting points should be equivalent.
3. The results obtained for the same sample of examinees using two different cloze tests should be equivalent.

Fairly recently, and unfortunately only after cloze tests had already achieved the important position in language testing that they hold today, two major investigations into these questions were reported: Alderson in 1978 and Klein-Braley in 1981. Their results showed that many of the claims put forward in favour of cloze tests were exaggerated.

To investigate question 1 Klein-Braley coded 13 texts into 12 word classes and then used a computer to construct all 45 cloze tests for each text which could be produced using the deletion rates 5, 6, 7, 8, 9 and 10 and all possible starting points. Statistical tests calculated on the basis of the two super-ordinate categories, structure and content words, showed that nth word deletion was not equivalent to random deletion. Indeed, empirical investigations showed that the actual difficulty of tests constructed from the same text varied according to the proportion of structure words deleted.

Both researchers examined the second question. Klein-Braley defined equivalence in terms of mean item difficulty and reliability coefficients. The results are shown in Table 4. For the same text difficulty varies between .36 and .65 (exact scoring) and reliability from .46 to .73 (exact scoring). For acceptable scoring the variability of difficulty is smaller (.73 to .87) but the reliability coefficients range from .38 to .62.

Table 4
Same text, different deletion rates

	Deletion	Mean item difficulty	Reliability
exact	8	.36	.46
	9.1	.56	.72
	9.3	.65	.73
	9.4	.61	.71
	10	.59	.62
accept.	8	.73	.60
	9.1	.79	.61
	9.3	.87	.38
	9.4	.83	.62
	10	.79	.60

Alderson defined equivalence in terms of validity coefficients: the correlations between the various cloze tests and the ELBA test. Again considerable differences were revealed. His easy text showed a correlation with ELBA of .59 at deletion rate 6 but of .70 for deletion rate 8; his medium text correlated .57 for deletion rate 6 but .86 for deletion rate 10.

In order to investigate question 3 Klein-Braley administered 6 different pairs of cloze tests to sample groups of students at the University of Duisburg (Table 5).

Table 5
Correlations between pairs of cloze tests
Administered to the same samples

	Exact scoring	Acceptable scoring
1 & 2	.55	.51
3 & 4	.39	.36
5 & 6	.48	.59
7 & 8	.58	.52
9 & 10	.70	.74
11 & 12	.64	.36

For acceptable scoring the correlations range from .36 to .75; for exact scoring from .39 to .70 (cf. Klein-Braley 1983:224).

These results, which represent only a very small part of the empirical findings of both researchers, offer no justification for the

notion that cloze tests are equivalent to each other. But the theory states that if cloze tests are operationalizations of reduced redundancy then they must be equivalent.

4 Conclusion

At the end of this review of tests of reduced redundancy where do we stand? We have a theory which states that one way to assess an individual's competence in a language, his internalized rule system, is to present him with a text where the redundancy has been reduced. In this way he is forced to activate all he knows about the language in order to restore the missing portions of the text. The better his competence, the more easily he will be able to do this. We can therefore draw conclusions about his competence from his performance on such tests.

On the other hand we have seen that out of three attempts to operationalise this theory in the form of tests, not one has proved satisfactory.

As Cronbach and Meehl (1955) point out:

The investigator whose prediction and data are discordant must make strategic decisions. His result can be interpreted in three ways:

1. The test does not measure the construct variable.
2. The theoretical network which generated the hypothesis is incorrect.
3. The experimental design failed to test the hypothesis correctly.

On other words, one can reject the theory together with the test. Or one can retain the theory and try to find a new and better operationalization. In Duisburg we still found the original theory convincing, and we therefore chose to use the third strategy. In 1981 Raatz and Klein-Braley first presented the C-Test to an IUS Symposium, and Raatz discusses our test of reduced redundancy in his paper (see Raatz, this volume).

Notes

- 1 It had been criticized as early as 1973 by Johansson, who particularly emphasised that the scoring technique was, in his opinion, unsatisfactory. His own noise test did not seem to function very well, but it is difficult to assess his results against those of Spolsky and his team since Johansson changed many of the determinants of the situation: he used lower signal/noise ratios (which theoretically ought to increase the difficulty) while at the same time constructing sentences deliberately very high in semantic and/or syntactic redundancy (which presumably reduce the difficulty). He also changed the scoring system. (cf. Johansson, 1973a).

- 2 It is difficult to see how dictation in the traditional form can be regarded as a test of reduced redundancy, although it is, of course, one of the major techniques advocated as a test of general proficiency.

Bibliography

- Alderson, J.C. 1978. A study of the cloze procedure with native and non-native speakers of English. Unpublished PhD Dissertation, University of Edinburgh.
- 1983. "The cloze procedure and proficiency." In Oller, J.W.Jr. (ed) 1983. Issues in language testing research. Rowley, Mass: Newbury House. 205-217.
- Carroll, J.B. 1961. "Fundamental considerations in testing for English language proficiency of foreign students." In Testing the English proficiency of foreign students. Washington D.C.: Center for Applied Linguistics. 30-40.
- , Carton, A.S. and Wilds, C.P. 1959. An investigation of cloze items in the measurement of achievement in foreign languages. College Entrance Examination Board Research and Development Report. Cambridge, Mass.: Laboratory for Research in Instruction, Harvard University.
- Chomsky, N. 1965. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press.
- Cronbach, L.J. and Meehl, P.E. 1955. "Construct validity in psychological tests." Psychological Bulletin 52: 281-302.
- Darnell, D.K. 1968. The development of an English language proficiency test of foreign students using a clozentropy procedure. Boulder, Co.: Department of Speech and Drama, University of Colorado.
- Gaies, S.L., Gradman, H.L. and Spolsky, B. 1977. "Toward the measurement of functional proficiency: contextualization of the noise test." TESOL Quarterly 11,1: 51-57.
- Gradman, H.L. and Spolsky, B. 1975. "Reduced redundancy testing: a progress report." In Jones, R.L. and Spolsky, B. (eds) 1975. Testing language proficiency. Arlington, Va.: Center for Applied Linguistics. 59-70.
- Johansson, S. 1973a. "An evaluation of the noise test: a method for testing overall second language proficiency by perception under masking noise." IRAL 11,2: 107-132.
- 1973b. Partial dictation as a test of foreign language proficiency. Swedish-English contrastive studies 3. Lund University, Dept. of English.

- Klein-Braley, C. 1981. Empirical investigations of cloze tests. An examination of the validity of cloze tests as tests of general language proficiency in English for German university students. PhD Dissertation, University of Duisburg.
- 1983. "A cloze is a cloze is a question." In Oller, J.W.Jr. (ed) 1983. Issues in language testing research. Rowley, Mass.: Newbury House. 218-228.
- Lado, R. 1961. Language testing. London: Longman.
- Miller, G.A. ²1970. The psychology of communication. Harmondsworth: Penguin.
- Oller, J.W.Jr. 1971. "Dictation as a device for testing foreign language proficiency." ELT 25: 254-259.
- 1972. "Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language." Modern Language Journal 56,2: 151-158.
 - 1979. Language tests at school. London: Longman.
 - and Conrad, C. 1971. "The cloze technique and ESL proficiency." Language Learning 21: 183-195.
- Shannon, C.E. and Weaver, W. 1949. The mathematical theory of communication. Urbana: University of Illinois Press.
- Spolsky, B. 1973. "What does it mean to know a language; or how do you get someone to perform his competence?" In Oller, J.W.Jr. and Richards, J.C. (eds) 1973. Focus on the learner. Rowley, Mass.: Newbury House. 164-176.
- , Bengt, S.M., Sako, E.W. and Aterburn, C. 1968. "Preliminary studies in the development of techniques for testing overall second language proficiency." In Upshur, J.A. and Fata, J. (eds) 1968. Problems in foreign language testing. Language Learning Special Issue No. 3: 79-103.
- Taylor, W.L. 1953. "Cloze procedure - a new tool for measuring readability." Journalism Quarterly 30: 414-438.
- Valette, R. 1967. Modern language testing. New York: Harcourt Brace and World.
- Winograd, T. 1984. "Software für Sprachverarbeitung." Spektrum der Wissenschaften. Nov. 1984: 88-132.

TESTS OF REDUCED REDUNDANCY - THE C-TEST, A PRACTICAL EXAMPLE

Ulrich Raatz (University of Duisburg)

1. Criticism of cloze tests

Cloze tests are unsatisfactory operationalizations of the concept of reduced redundancy, and therefore also of the theory of expectancy grammar, as Alderson (1978) and Klein-Braley (1981) have shown. In addition to the criticism mentioned, they also pointed out:

1. The deletion rates used in classical cloze tests are too high. If 50 items are usually demanded in order to ensure an acceptable level of reliability then extremely long texts are necessary. In addition the deleted words are not a representative sample of different word classes of the text. This means that cloze tests developed from the same text but with different starting points and deletion rates differ in terms of difficulty, reliability and validity, as Klein-Braley showed quite clearly in her study.
2. If one uses only one text, as is standard practice in cloze tests, one cannot assume that this is a representative sample of the language.
3. The two different scoring methods for cloze tests are problematical. If the exact method is used cloze tests are too difficult even for competent native speakers. But the acceptable scoring method is much less objective and extremely time-consuming. This is true especially for the cloze-entropy scoring technique (Darnell 1970).
4. The basic requirement that adult educated native speakers of a language should score full marks on a tests intended for non-natives is not fulfilled by any cloze tests.

This criticism led to the establishment of criteria which an improved procedure ought to fulfill, although these criteria were pure theory and seemed utopian. They were:

1. The new test should be much shorter, but at the same time it should have at least 100 items.
2. The deletion rate and the starting point for deletions should be fixed, and no longer a matter of personal preference.
3. The words affected by the deletions should be a genuinely representative sample of the elements of the text.

4. Examinees with special knowledge should not be favoured by specific texts, in other words the new test ought to contain a number of different texts.
5. Only exact scoring should be possible so as to ensure objectivity.
6. Native speakers ought to be able to score 100% on the test.

In addition the test, of course, had to be reliable and valid and very easy to develop.

These requirements seem impossible to fulfil if the classical cloze principle is retained. The principle must be modified. One modification is the C-Principle. The tests constructed using this principle are a very much better operationalization of the underlying theory than classical cloze tests.

2. C-Principle and C-Tests

It is obviously the deletion technique which leads to the difficulties encountered with the classical cloze test. This was the reason that we decided to start here in our attempts to improve the tests. We developed a new deletion technique which we called the C-Principle. Deletions are not performed at the sentence level but at the word level. We no longer remove whole words from the text; we "damage" parts of words.

How does the C-Principle work? In short, carefully selected, preferably authentic texts we delete the second half of every second word, beginning with the second word of the second sentence. If the word has an odd number of letters then half a letter more is deleted. If the word only has one letter, as, for instance, in the English words "a" and "I", then this word is ignored in the counting. A number of texts which have been mistreated in this way, usually between 4 and 6, are put together to form a "C-Test". The examinee is confronted with these texts in the form of a test paper and is asked to restore the deleted elements. He is given a reasonable amount of time to perform this task. Only entirely correct restorations are counted as right.

The C-Tests which have been developed according to this principle have the following substantial advantages over cloze tests:

1. Many more items are possible with much shorter texts. A classical cloze test using a 5% word deletion rate would have to be at least 500 words long in order to contain 100 items. A C-Test consisting of 5 texts with 20 half deleted words would be only approximately half as long.
2. C-Test scoring is "exact" and objective because there is almost always only one possible solution. In a few exceptional cases there are a small number of different solutions

(e.g. this and that). In such cases all solutions are accepted.

3. The scoring of a C-Test is very quick and easy for a native speaker. He or she only needs to read through the text once, and normally needs no key.
4. C-Tests are very "easy" for native speakers, because the first half of the word is left standing. In a number of different pilot studies it has been shown that native speakers were normally able to restore over 95% of the damaged words. On the other hand, someone who does not understand the language normally scores zero on the test.
5. Since every second word is "damaged" the probability of obtaining a representative sample of all the word classes in the text is very much higher. This was examined in a special investigation, which was aimed at revealing whether deletion of every second word in a 200 word text produced a representative sample of all the words in the text. Two students were asked to examine 40 English and 40 German texts, each with 200 words, chosen at random to vary as much as possible so far as style and content were concerned. They coded each text into the categories: nouns, adjectives, verbs, adverbs, prepositions, articles and remainder. Then they deleted every second word and compared the distribution of the deleted elements to that of the elements in the original texts. The results showed that for every one of the 80 texts the deleted samples are representative for the entire text.
6. Because a C-Test consists of a number of different neutral texts the sampling of content classes is better. Test subjects who happen to have special knowledge in certain areas have no substantial advantage over the other subjects.

C-Tests are an operationalization of the theory of reduced redundancy. This one can see clearly if one looks at the overview on page . In contrast to the other procedures a C-Test consists of 5 to 6 short texts. Deletion takes place at word level. The channel is visual, as in the cloze test, and the examinee behaviour is also similar to that demanded by the cloze test, although here one has to use the term gaps rather than blanks to indicate that only parts of words have to be restored.

There is however one mayor difference. Redundancy is present in language on a number of different levels: on the level of sentences, of words, and naturally on the level of the letters inside the word. This third type of redundancy can be used in the C-Test too, to help in finding the appropriate solution. The processes which take place inside the subject when he or she is working through a C-Test are currently under investigation.

Using the C-Principle, C-Tests have been developed in several different languages. A variety of subject groups have been tested, and a number of different questions have been investigated. The languages belong typologically to the analytic and synthetic classes: German, English, French, Spain, Turkish, Dutch and Ivrit. However no studies have been conducted so far with a genuinely agglutinating language such as Finnish.

The following figures show examples of parts of C-Tests in two different languages:

Figure 1

Part of a German C-Test

Gestern ist eine 76 Jahre alte Frau in Garath bei einem Unfall todlich verletzt worden. Die Fußga_____ hatte i_____ Höhe d_____ Stettiner Str_____ die Frank_____ Straße überd_____ wölien u_____ war da_____ von ei_____ Auto angef_____ worden. D_____ Ampelanlage a_____ dieser Straß_____ war z_____ Zeitpunkt d_____ Unfalls abges_____ und wu_____ repariert. I_____ Krankenhaus er_____ die Frau ihren schweren Kopfverletzungen.

Figure 2

Part of an English C-Test

Pollution is one of the big problems in the world today. Towns a_____ cities a_____ growing, indu_____ is gro_____ and t_____ population o_____ the wo_____ is gro_____. Almost every_____ causes poli_____ in so_____ way o_____ another. T_____ air i_____ filled w_____ fumes fr_____ factories a_____ vehicles, a_____ there i_____ noise fr_____ aeroplanes a_____ machines. Riv_____, lakes a_____ seas a_____ polluted b_____ factories and by sewage from our homes.

In the English versions each text has 25 deletions, because English words tend to be shorter on average. This phenomenon had been adjusted for by raising the number of deletions.

3. How is a C-Test developed?

First of all one looks for a number of suitable short texts in the appropriate language which are complete in themselves. The texts should preferably be authentic, and neutral in their subject matter. In developing tests for schoolchildren in their native language we have found school text books very useful, but of course these must be ones which are not known to the target group. Readers in particular are a rich source for suitable texts. Texts for students or adult groups were obtained from reference books and other nonfiction sources, or newspapers.

One then deletes 20 or 25 half words from the texts. The deletions normally begin in the second word of the second sentence. At the end the text runs on until it reaches a natural conclusion.

Four to six texts which have been treated like this are assembled to form a C-Test. Just as in the construction of any test it is a good idea to order the texts in ascending difficulty.

The time allowed for working through the test should be generous. We have found that in general 5 to 7 minutes are needed for each text in the test.

To score the text the scorer reads through the text and marks the correct solutions. The total score on the test is the sum of the part scores. In those cases where items have more than one possible solution all these solutions are counted as correct. Our experience has shown that this method of scoring is objective and fairly quick to carry out.

Problems can arise if C-Tests are to be developed for foreign language learners in the early stages of learning. Subject groups of this kind have only a very limited vocabulary, and this is defined by the textbook they are using. The same applies to their repertoire of structures. If C-Tests for such groups were constructed from authentic texts they would be both too difficult and unfair. In such cases we recommend the analysis of the vocabulary and structures used in the textbook, followed by the construction of quasi-authentic texts by native speakers, using only those words and structures which are known to the pupils. The other alternative is to use texts taken from textbook series which are not being used by the target group, but which use the same linguistic material. Here, of course, modifications can be necessary.

Item analysis in the classical sense, i.e. on the deletion level, which usually takes place after the first trial run and which is used to improve the test, is neither possible nor desirable with the C-Test. This is because the texts used are authentic, and it would not be possible to delete unsuitable items without

destroying both the sense of the text and the deletion principle. Furthermore the items cannot be statistically independent of each other, since they are all embedded in the same linguistic context. The only way a C-Test can be improved is by deleting or changing an entire text.

This means that item analysis for C-Test must be performed on the text or super-item level. In this analysis the difficulty of a text (superitem) can be defined as the percentage of correct solutions, the discrimination as the part-residual correlation of the super-item with the total score, and the reliability as Alpha-reliability according to Cronbach. In order to examine the question of whether all the texts in a C-Test measure the same overall proficiency factor analysis is not a suitable technique (Raatz, 1982). A better method is the use of the CLA model developed by Moosbrugger and Müller (1982) which is similar to the well-known Rasch model.

4. Investigations into the reliability and validity of C-Tests

When a new approach to the measurement of general language proficiency is proposed, an approach which claims to be suitable for a variety of languages a number of different target groups including L1, L2 and LF learners, and to be useful for a variety of purposes, then it is essential to perform a large number of empirical investigations, varying all possible conditions. The languages, the groups, the age levels and the problems must all be investigated to discover whether the C-Principle really produces reliable, valid and useful tests. If it can be shown that one particular C-Test, for instance a German C-Test, developed for a specific target group and to be used for a specific purpose, is sufficiently reliable and valid, this is very gratifying for the researchers, but it says nothing about the generalisability of the C-Principle.

For this reason we have spent the last few years performing a large number of investigations into the performance of C-Tests for a variety of languages, groups, age levels and problems. Other colleagues, both inside and outside of Germany, have performed their own investigations.

In virtually all cases reliability was estimated using Cronbach's Alpha. It is common knowledge that this formula underestimates the measurement accuracy of a test procedure.

In order to answer the question of what a C-Test actually measures, concurrent and predictive validity were investigated, in some cases factorial and construct validity were examined. Sometimes our external criteria were school grades in the appropriate language, sometimes they were ratings of the general level of proficiency attained given by teachers, in some cases we had the results of other test batteries which had been administered at the same time. In the investigations of factorial validity we used among other things the results of other language tests and intelligence tests.

4.1. C-Tests with L1 groups

Table 1 shows the results of 6 major investigations into reliability and validity with L1 groups. These studies were done by us with the help of students at our university and local teachers. The language of the C-Test was German in the first four investigations, English in the fifth and Turkish in the sixth. The pupils came from the second to eleventh grades. The third column shows the number of subjects involved in each investigation, the fourth the reliability coefficients. In those cases where a number of coefficients were obtained the lowest and highest are given, or the median of all reliability coefficients. The fifth column shows the concurrent validity. The criteria are school grades and teacher ratings. In one case we used pupil self assessments. In the second investigation the criterion was determined six month after the test administration.

Table 1
Reliability and validity of C-Tests
for different groups and languages

Language	Sample	L1 Groups		
		N	r_{tt}	r_{tc}
German	schoolchildren 3rd - 8th grade	716	.76 - .88	.47 - .80 school grades
German	schoolchildren 2nd grade	117	.83	.63 (med) ratings (half year later)
German	schoolchildren 5th grade	75	.78 - .81	.57 - .65 school grades factorial
German	schoolchildren 6th - 11th grade	402	.78 (med)	.40 (med) school grades
English	schoolchildren 10 - 15 years	160	.66 - .92	.73 - .88 ratings/self assessment
Turkish	schoolchildren 17 years	13	.90	.45 ratings

At the last two columns show, the reliability coefficients range between acceptable and high, the validity coefficients are most satisfactory.

4.2. C-Tests with L2 groups

Table 2 shows the results of six investigations into general language proficiency in German as a second language (5 investigations) and Hebrew as a second language. Studies 1 to 5 were in the main conducted by Edgar Süßmilch from Düsseldorf in cooperation with us. The investigations into adult learners of Hebrew as a second language were performed by Andrew Cohen in Israel. The children involved in studies 1 to 3 are the children of immigrant workers living in Germany. Again the third column shows the size of the sample involved, the fourth the reliability coefficients, some of which are remarkably high, and the fifth column shows an estimate

Table 2
Reliability and validity of C-Tests
for different groups and languages

Language	Sample	L2 Groups		
		N	r_{tt}	r_{tc}
German	Turkish children 2nd - 4th grade	227	.85 - .94	.41 - .88 school grades and ratings
German	Greek children 3rd - 4th grade	305	.82 - .92	.58 - .90 school grades and ratings
German	other nations	87	.83 - .88	.73 school grades and ratings
German	Turkish pupils 17 years	15	.86	.72 school grades
German	Turkish girls	37	.76	.57 awo test (3 subtests)
Hebrew	adults	62	.93 - .95	.69 - .87 (lang. tests) .72 - .92 (ratings)

of concurrent validity. Again the criteria are ratings and school grades with the exception of the last two studies where other language tests were used. Here again the results are very positive.

4.3. C-Tests with LF groups

Table 3 shows the results of 7 investigations which we either conducted ourselves or which were reported to us.

The first investigation shows the results obtained for a multinational group of young people and adults learning German at a language school. The criterion for this investigations was the EUROTEST, a discrete-point item test with 4 subtests used for placement in the language school. Reliability is very high, and so is the agreement between the C-Test and the discrete-point item test.

Table 3
Reliability and validity of C-Tests
for different groups and languages

Language	Sample	FL Groups		
		N	r _{tt}	r _{tc}
German	Foreign students	252	.92 - .93	.82 - .85 Eurotest (4 Subtests)
English	German school- children 6th grade	108	.81	.60 school grades
English	German students	438	.77 - .91	.62 - .90 Delta Test (6 Subt)
English	Finnish students	51	.87	
English	Belgian students	192	.77	
French	German students	79	.97	.87 Bochum French Plac. Test (7 Subt)
Spanish	German adults (Further educ)	24	.92 - .94	.90 - .95 ratings

The second investigation shows the results obtained for 108 German pupils learning English in the second year of secondary school. For this study a C-Test consisting of pseudo-authentic texts using only the vocabulary and structures known to the pupils and written by native speakers of English was developed.

The subjects in the third study are students of English at the University of Duisburg who have to take a placement test before being assigned to the appropriate language courses. They took a C-Test together with the DELTA test, a discrete-point item test with 6 subtests.

The next two investigations were performed with Finnish students by Hilka von Essen and with Belgian students by Bea von Beckhoven. We did the computing. For these investigations we have no validity coefficients.

The last investigation was performed by Grotjahn and Stemmer in Bochum. The subjects were 79 beginners in the French department taking the Bochum placement test in French, again a discrete-point item test, which was administered together with the C-Test.

The Spanish test, which we developed ourselves, was used in a pilot study. Here the subjects were adults learning Spanish at the local evening institute in their spare time.

As Table 3 shows, the results of the investigations with LF groups reveal perhaps even more clearly than the ones carried out with L1 and L2 groups that C-Tests in different languages and for different target groups are reliable, at times very highly reliable tests, and that in all cases they have high concurrent or predictive validity. We find it particularly interesting that C-Tests used for placement purposes correlate so highly with the results obtained with discrete-point item tests. This could lead to interesting developments in placement procedures.

4.4. Investigations into construct validity

In our recent work with the C-Test we have been examining various aspects of construct validation. The C-Test is based on the same fundamental theory as cloze tests. We posit an internalized grammar, or rule system, which is in the course of evolving in the learner. At any given point of measurement this system is self-consistent. As the learner progresses, he or she revises and adds to the system so that it becomes more and more similar to that of the adult educated native speaker of a language. By using a test of reduced redundancy in an acceptable operationalization we can force the learner to activate the internalized grammar in order to restore the missing elements in the text. The higher the score on the test, the more accurate, we conclude, is the learner's grammar, i.e. we draw conclusions from the learner's performance about his or her competence.

We tried to set up a number of predictions which followed from this theory and to examine whether they could be empirically confirmed.

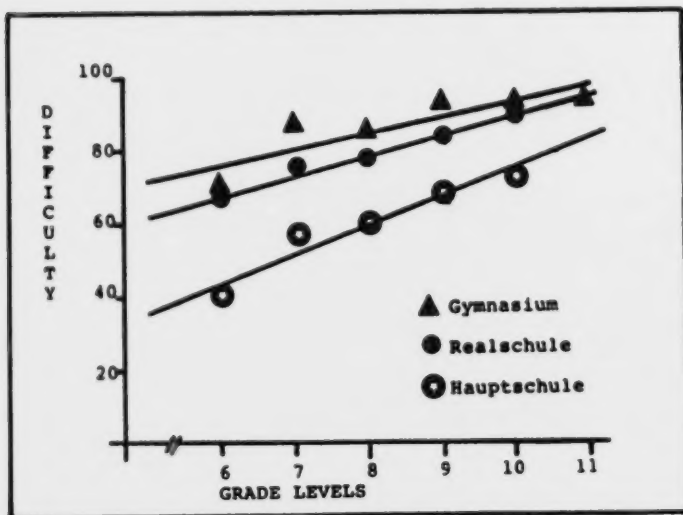
Our first hypothesis involved the basic question of whether C-Test performance was similar to performance in other language tasks. This, of course, is simple concurrent validation, and the results were positive.

We then turned our attention to age. For L1 groups, we thought, test scores ought to improve with increasing age. We could show, that relationship turned out to be linear (see Fig. 3).

The same study showed that the mean performance of pupils in the three different types of secondary schools in German differed, and that the results were in the direction predicted.

Figure 3

Empirical text difficulty levels and regression on age for pupils at different types of German secondary schools



Our next question was the rôle played by intelligence. We expected a moderate positive relationship between IQ scores and C-Test performance. We were able to show that this was the case, and also that C-Test scores in a factorial study clustered with other language scores (grades and a diagnostic test in L1, grades in LF) which indicated that a general non-language-specific verbal ability could be involved. The second factor was an intelligence factor with a speed component. (Ratz, 1984)

Klein-Braley investigated the hypothesis that, in terms of C-Test processing, texts had an inherent constant level of difficulty so that the observed C-Test score for any individual was a function of text difficulty and subject ability. It ought therefore to be possible to discover text characteristics which would enable us to predict test difficulty for specific groups. Again the results were positive. (Klein-Braley, 1984)

I was then able to show by use of a special latent trait model that this function for any given subject was an additive one. (Ratz, 1985)

We also investigated the question of the transferability of the C-Principle to other languages. We began our investigations with English and German. Grotjahn added French. I could show that the principle worked for Spanish. These are all Indo-European languages, and so we wondered whether it might be some particular feature of this language family which was responsible for the results. We felt that the validity of the principle received its confirmation when we were informed by Cohen (1985) that a satisfactory test had been constructed in Hebrew, and by Süßmilch's development of a test in Turkish.

So although the construct validation of C-Tests is still continuing - we are particularly interested in some work going on in Bochum (Stemmer, 1984) where C-Test processing strategies are under investigation - we feel at this point that we can claim construct validity for the C-Test.

5. Where can C-Tests be useful?

According to the underlying theory C-Tests provide reliable and valid information about the general level of proficiency in the native language or in a foreign language; they enable us to draw conclusions about the state of development an individual's competence in a language has reached from the assessment of his or her performance. However they provide no detailed information about strengths and weaknesses in individual areas of the language. So they tell us nothing about specific features of grammar, or syntax, or vocabulary, or spelling, and they offer no direct evidence about performance in listening and speaking. For this reasons the possible uses of C-Tests are limited.

For schools, universities and institutes of adult and further education we see the following four main areas where C-Tests could provide useful information:

1. C-Tests can be used at the beginning of a school year, a university term or a teaching unit as placement tests for determining which of a number of possible language courses is most suitable for an individual learner. They could also be used to select those students with very high proficiency levels in foreign languages.
2. It has proved difficult to develop suitable test procedures for the children of immigrant workers learning a second language. Such tests are needed to decide whether children should be given special tuition in the second language or whether they are able to attend normal classes together with L1 speakers of the language. Similar problems can also arise for adults, for instance in Israel.
3. C-Tests can be used as global diagnostic instruments of learning difficulties in language, both in the native language (dyslexia) and in the foreign language. They would, of course, need to be backed up by other diagnostic procedures.
4. Since the same C-Tests can be used with L1, L2 and F1 learners they can be used in comparative studies as research instruments, for instance to investigate questions of language development and language learning.

6. Conclusions

It appears that the C-Test does everything that the cloze test promised. Speaking as the researchers who conducted the vast majority of the investigations reported in this paper we can state that we are satisfied with the performance of the tests for all the languages and for all the groups with which they have been used. Nevertheless we must stress that a C-Test is not guaranteed to be automatically reliable and valid. It seems probable that, provided the C-Test is suitable in its level of difficulty for the target group envisaged, it will, in the majority of cases, produce valid results.

At the same time, before C-Tests are used in any institution for real decisions which would affect pupils' or students' lives, a certain amount of statistical evaluation is essential.

What must not happen is that uncritical users begin to abuse the C-Principle and the C-Tests by failing to perform the necessary statistical checking.

Bibliography

- Alderson, J.C. 1978. The effect of certain methodological variables on cloze test performance and its implication for the use of the cloze procedure in EFL testing. Paper presented at the 5th International Congress of Applied Linguistics, Montreal, Canada.
- Cohen, A.D., Segal, M. and Weiss, R. 1985. "The C-Test in Hebrew". In Klein-Braley, C. and Raatz, U. (eds). 1985. C-Tests in der Praxis. Bochum: AKS.
- Darnell, D.K. 1970. Clozentropy: a procedure for testing English language proficiency of foreign students. Speech Monographs 37, 1: 36-46.
- Grotjahn, J. and Stemmer, B. 19 . "On the development and evaluation of a C-Test for French". In Klein-Braley, C. and Raatz, U. (eds). 1985. C-Tests in der Praxis. Bochum: AKS.
- Klein-Braley, C. 1981. Empirical investigation of cloze tests. An examination of the validity of cloze tests as tests of general language proficiency for German university students. Universität Duisburg: Inauguraldissertation.
- 1984. "Advance prediction of difficulty with C-Tests". In Culhane, T., Klein-Braley, C. and Stevenson, D.K. (eds). 1984. Practice and problems in language testing 7. Colchester: University of Essex.
- Moosbrugger, H. and Müller, H. 1982. "A Classical Latent Additive Test Model (CLA Model)". The German Journal of Psychology 6: 145-149.
- Raatz, U. 1982. "Language theory and factor analysis". In Lutjeharms, M. and Culhane, T. 1982. Practice and problems in language testing 3. Proceedings of the Third International Language Testing Symposium of the IUS. Brüssel: Vrije Universiteit.
- "The factorial validity of C-Tests". In Culhane, T., Klein-Braley, C. and Stevenson, D.K. (eds). 1984. Practice and problems in language testing 7. Colchester: University of Essex.
 - "Better theory for better tests". To appear in Language Testing 1985.

TESTING LANGUAGE SKILLS IN THE NORWEGIAN COMPREHENSIVE SCHOOL

Lars Sigfred Evensen (University of Trondheim)

1 Introduction

Finland is at a stage where streaming in FLT is on its way out of the system and mixed-ability teaching and testing on its way in. Such a decision is going to have far-reaching effects on both teaching and testing. It is against this background that I would like to present selected data from the Norwegian system.

In Norway a curriculum based on the concept of mixed-ability teaching has been in effect since 1974. It grew out of a long period of pedagogical and even political struggle. In fact, the more fundamental questions behind the mixed-ability concept (eg. should we have one school for all?) have been central in Norwegian debate for more than one hundred years (Telhaug 1970:10). The reason for this is that one's choices on such issues will have effects that are not only linguistic and pedagogical, but also social. Education gives access to so many other things in society.

In the context of mixed-ability teaching testing is but one aspect. However, major pedagogical questions like individualization have already been topics for separate conferences and seminars in Finland (cf. eg. seminar reports No.3 and No.11 from Nordiska språk- och informationscentret, Helsinki). Norwegian points of view have been presented by Frydenberg 1981a, 1981b and 1981c, by Stølen 1984a, 1984b and by Tangen 1984. Thus in my present talk it is natural to concentrate on questions of evaluation and assessment in English as a foreign language (EFL) in general and the exam system in particular.

As an applied linguist I feel that it is very important to supplement information about a system with info from the people who are actually working within the system. It is, after all, the wearer of the school, so to speak, who knows where it hurts most. For this reason I shall present selected results from a national survey carried out in 1981. The aim of this survey was to find out what a statistically representative national sample of students and teachers in grades 8-11 (student ages appr. 15-18) of Norwegian lower and upper secondary school perceive to be their problems in LL/LT in Norwegian as a first language and EFL (Evensen 1983).

In the survey evaluation stands as one out of a considerable number of problem areas. In the context of this conference I feel that it may be just as relevant to hear about such problems as to hear about any solutions that we might have to offer.

2 The Norwegian EFL curriculum and evaluation system

Teaching in Norwegian comprehensive school has since 1974 been based on the "Mønsterplan for grunnskolen" (Curriculum for primary and lower secondary education). The main educational innovation introduced by this plan was the step away from a fixed curriculum for all to a flexible teaching guide, where individual schools and teachers (at least in principle) are given considerable freedom to choose for themselves both what to teach and how to teach it. The testing system, however, is still largely centralized.

In the EFL teaching guide a major change consisted of moving away from explicit knowledge of language structure and metalanguage to an explicit emphasis on skills and a less explicit emphasis on language as a means of communication. The goal of EFL is stated in the following three parts (my translation):

- The teaching of English is to aim at
- giving the students practical language skill which may give them increased ability to establish contact, orally and in writing
- forming as firm a basis as possible for continued learning of the subject
- establishing such attitudes toward the language that the students develop an interest in increasing their skills and get an impulse to use the language."

Based on considerations of communicative needs, oral skills are given more emphasis than written skills. Furthermore, of the two written skills reading seems to be given priority before writing, which is partly given a secondary role (Simensen 1983):

- "The skill of written expression is also to be practiced, both as a support of training of the oral skills and as a means of expression in itself".

Given the fact that the teaching guide is based on general goals rather than specific objectives, with considerable local freedom within the system, one may suspect that teachers have problems in knowing what to evaluate, which criteria to use. In the national survey referred to, the teacher questionnaires contained the following question:

Do you think it is easy or difficult to know what to correct in this grade? (Put a ring around the figure below the appropriate response alternative).

very diffic.	quite diffic.	medium	quite easy	very easy	don't know
5	4	3	2	1	6

Figure 1. Translated item from EFL teacher's questionnaire.

The answers of 191 EFL teachers in grade 9 are displayed in Table 1 below. In the table answers are given as percentages for each response alternative. The frequencies are weighted to counterbalance biases introduced by disproportionate sampling design (cf. Evensen 1982a). Furthermore the frequencies are adjusted so that those respondents who answered 'don't know' (1.0 per cent of the total) or left the item unanswered (3.1 per cent of the total) are kept out of the computations.

Table 1. Selecting criteria for correcting student performance. EFL teachers' perceptions of difficulty in grade 9. Weighted and adjusted frequencies (n=191).

response alternatives					central tendency and dispersion	
very diffic.	-----			very easy		
5	4	3	2	1	Md	s (1)
%	%	%	%	%		
02.9	13.5	49.5	27.3	06.8	2.821	.868

16.4 per cent, or about a sixth, of the teachers feel that knowing what to correct is very difficult or quite difficult. For these teachers the choice of evaluation criteria may be said to constitute a problem. For the great majority, however, the system seems to work quite well. If we use the response scale from five to one to compute an average (median) out of the answers given, the result is 2.821. This means that an "average answer" is a little toward the 'easy' end of the scale. The standard deviation adds the further information that there is some disagreement among teachers on the question; there is some degree of dispersion in the answers given.

The most reasonable conclusion one can draw from the data seems to be that, after seven years of experience with mixed-ability teaching, most teachers are in fact quite satisfied with the freedom (or burden) of choice built into the system. The table, in other words, seems

to show that local freedom is not as big a problem as one might have expected it to be for teachers with respect to knowing what to correct.

We may go on to ask what the students think of the teachers' evaluation. The student questionnaires contained an item of the type presented above where the students were asked to evaluate teacher demands, as demonstrated in teacher marking, on an ordinal scale going from 'much too high' to 'much too low'. The answers are presented in Table 2.

Table 2. Marking. EFL students' perceptions of teacher demands in grade 9. Weighted and adjusted frequencies (n=410).

response alternatives					central tendency and dispersion	
demands are---						
much too great -----			much too small		Md	s
5	4	3	2	1		
%	%	%	%	%		
12.7	30.9	49.7	06.8	00.0	3.371	.801

It seems that the students are not totally satisfied with the state of affairs. 43.6 per cent feel that teacher demands are 'much too great' or 'a little too great'. On the other hand nobody seems to feel that the teachers are much too lenient. 6.8 per cent think that the teachers are a little too lenient. The median reflects this distribution. A result of 3.371 is toward the 'too great' part of the scale. Still, the median is so close to the neutral mid category that this problem may not in any sense be said to be an important one.

3 The Norwegian EFL exam

In this section I shall present some information about the Norwegian EFL exam at the end of comprehensive school. In the presentation I shall focus on the correspondence between the goals and the exam system. I shall also consider potential backwash effects on everyday teaching. It is a well-known fact that the exam system influences practical teaching. Simensen (1983:217) has formulated this principle very strongly; if there is a discrepancy between goals and evaluation, evaluation usually wins.

I have chosen to use teacher perceptions of correspondence as my starting point. In the questionnaire teachers were asked to give their views as to the degree of correspondence between the goals and the EFL exam. I find the results very interesting.

Table 3. Teacher perceptions of the correspondence between the official goals and the EFL exam at the end of comprehensive school. Weighted and adjusted frequencies (n=191).

response alternatives					central tendency and dispersion	
very good corr.				very bad corr.	Md	s
5	4	3	2	1		
%	%	%	%	%		
00.0	03.6	49.8	36.8	09.9	2.568	.722

It seems clear that we deal here with a real problem of some intensity. None of the teachers think that the correspondence is 'very good', and only 3.6 per cent feel that it is 'quite good'. On the other hand almost ten per cent think that the correspondence is 'very bad'. The median is as low as 2.568 and the low standard deviation (.722) indicates a higher degree of consensus among the teachers on this question than on previous ones. In an effort to understand this problem I shall first present some background information about the exam system and then supplement this material with more detailed survey data.

English is the only foreign language with a final written exam in Norway. The test takes place at the end of ninth grade and is allotted four hours (+ fifteen minutes for practicalities). It officially consists of at least four out of a list of eight subtests presented in the Handbook for the school, part III (Håndbok for skolen, del III):

- comprehension questions on a given text
- fill-in-the-blank (including close)
- completion of a text where the beginning or end has been removed
- free comments on a text
- essay based on key words
- picture-based essay
- essay on a given topic

According to the regulations each student sits for at least one of the following subjects-

- Norwegian as a first language
- Mathematics
- EFL

In practice each student sits for only one of these, which means that all students in approximately one third of the country's nineteen counties sit for the EFL exam every year (cf. Table 5 below). The particular counties are selected anew each year according to a confidential system.

Each student's performance is evaluated by a team of two external examiners each evaluating about 150 student performances. Examiners work independently at first, then change student answers and at the end discuss borderline cases.

Since 1982 the guiding principle behind evaluation criteria and marking has been one called "adapted goal-referenced testing" (cf. Norsk Skole Nos. 10/11 and 12/13 1982 and 4 1984). This principle may be seen as an attempt to combine goal-referenced, criterion-referenced and norm-referenced testing.

The main basis for marking is the examiners' overall impression of each student's performance. In doing this the examiners are told to relate their judgement not just to linguistic aspects, but also to the overall goals of the school system, as these are stated in the teaching guide. Apart from a general description of intentions the examiners are not given many details as to how to go about implementing the principle in actual marking.

More detailed evaluation criteria are distributed every year to the examiners and later published in the series *Evaluering i grunnskolen* (Evaluation in the comprehensive school) issued by The National Council for Primary and Lower-secondary Education.

For the more closed subtests in the written exam, some of which are discrete-point, a number of points are given for each subtest (implying a relative ranking of them as to importance) and a total sum for all subtests is suggested for each mark. In certain cases, a graded system is used for the allotment of points to each individual answer, where the top score is reserved for those answers that are both formally and functionally correct (eg. a question with do-periphrasis). A medium score is given for understandable but formally incorrect answers (eg. "How long do the journey takes?"), and no points are given for answers that are both formally and functionally incorrect. Thus communicative principles have to some extent entered into the criterion-referenced part of the marking.

For the essays several broad criteria are suggested. The essays in the upper two bands of the five-point marking scale (see below) should be "good" as to coherence (2), diction and vocabulary, having few errors of orthography and verb inflection. Furthermore the essays have to be in accordance with the wording of the task given.

The essays in the mid group are accepted with somewhat lower standard as to coherence, diction, vocabulary, errors of orthography and verb inflection; some naive formulations and a relatively short answer. The criteria for the low group are more indirect; a very short essay will not receive good marks and an otherwise medium essay should not be lowered more than one step in the marking if the student has not written the essay part of the exam. For the lowest mark the criterion is communicative; the essay would not be understood if performed orally.

By way of contrast, the oral exam consists of a combination taken out of a list of four subtests (Håndbok for skolen, del III):

- reading of a passage from a known text followed by conversation about the text from which the passage is taken
- free conversation about everyday topics
- reading of a short, unknown text followed by comprehension questions
- conversation about an unknown text read aloud by the teacher (the internal examiner)

Certain students from two counties are selected to sit for the oral exam in English every year. This means that the proportion of students sitting for the oral is much lower than the proportion sitting for the written exam (cf. Table 5 below).

There is one external examiner for each student, and each team of examiners evaluates the performance of some 20 students. In the marking regulations the emphasis is explicitly on the students' ability to carry on a conversation, which is more important than her/his pronunciation and intonation. The regulations do, here as well, have a communicative element.

Statistics as to the distribution of subtests are only available for written exams. In the following section I shall concentrate on the written exam, supplementing some statistics and survey data that are relevant also for the oral.

The norm-referenced part of the evaluation criteria is quite indirect. National standardized tests in the structuralist-psychometric tradition (Spolsky) have since 1962-63 been offered on an optional basis as one

element guiding the teachers' achievement marking. This system is used by 80-90 per cent of the schools (OMI report No. 20, p. 18). Although no norm is given for the national exam, it may be assumed that these tests have long-range effects on individual teacher marking both because of their long tradition and because of their widespread use (for further evidence, see Table 5 below). The weight of the test is said to equal one double-lesson school test. The marks are to be normally distributed (S is the top mark - "particularly good" - and Lg is the bottom mark - "little of positive value"):

S	M	G	Ng	Lg
++	+	/	-	--
4%	24%	48%	24%	4%

Let us look at the characteristics of the written EFL exam during the last five years. The distribution of subtests is presented in Table 4.

Table 4. Distribution of subtests in recent EFL written exams.

	1980	'81	'82	'83	'84
-open-ended comprehension questions	X	X	X	X	-
-true/false items	X	-	-	-	X
-text-based questions/answers	X	X	X	X	XX
-story/dialogue completions	-	-	-	-	X
-vocabulary test	X	-	X	XX	X
-cloze/ fill-in-the-blank	X	-	X	X	XX
-transformations and paraphrases	X	XX	-	-	-
-sentence completions	-	-	X	-	-
-text-based free comments	-	-	-	-	-
-expression of selected language functions	-	X	-	-	-
-letters/ postcards	X	-	X	X	-
-essay (1 out of 4-6 tasks)	X	X	X	X	X

The distribution of subtests clearly indicates that the average number of subtests is larger than four. Furthermore, several subtests that are not included in the official list appear in the actual exams (eg. transformations and sentence completion) whereas one subtest is not used at all in the period studied (free comments on a text). Two test types have dominated the exams in the period in the sense that they have been used every year - making questions or answers to a text and essay writing.

The frequency of the former subtest calls for a comment. The most likely explanation for this perhaps unexpected pattern is suggested in a critical study by Simensen (1979). On the basis of a longitudinal study of EFL curriculum development and exams she concluded that several aspects of the exam tradition might be questioned when related to the official curricula. One such aspect is the preponderance of items where the students are only trained to answer given questions when in real-life communication they also need to be able to ask questions. An inspection of the exam tasks in the period under study shows that it is in fact this aspect that is emphasized; this subtest in all cases following Simensen's report asks the student to make questions that are relevant to a given text or situation.

In Simensen's (1979) study the suspicion that some subtests may not discriminate well was also expressed. She furthermore suggested that the use of meta-language and discrete-point tests of grammar should be abolished. In the period under study it appears that there has been an increase in the number of subtests where two versions differing in difficulty (eg. in that one subtest is guided where a similar subtest is not or that the marking of one subtest takes accuracy into account where the other does not) are used.

In the period after Simensen's study no subtests have relied on metalinguistic knowledge. However to some extent discrete-point tests of grammar have still been used, but not since 1982. It may be justified to conclude that one positive aspect of the Norwegian system is its ability to take outside criticism into consideration.

One point that is not taken into consideration, however, is Simensen's criticism of the written language bias in the system. This bias demonstrates itself in the distribution of written and oral exams in the period under study. Table 5 presents both this distribution and the distribution of marks in the four-year period between 1980 and 1983 (the data from 1984 are not yet available).

Table 5. Marks, norms and exam results in EFL 1980-83.

Marks	S	M	G	Ng	Lg
	++	+	/	-	--
Norms	4%	24%	48%	24%	4%

Achievement marks:

1980 (n=62.648)	2.91	31.11	40.46	23.40	2.12
1981 (n=64.329)	3.26	30.87	40.72	23.02	2.12
1982 (n=65.895)	3.24	31.32	40.51	22.75	2.18
1983 (n=65.219)	3.18	31.72	40.70	22.35	2.05

Written exams:
(all students in
seven counties)

1980 (n=25.011 =38.2%)	3.40	22.73	43.51	25.75	4.61
1981 (n=26.975 =40.2%)	3.99	23.73	44.18	23.80	4.30
1982 (n=22.331 =35.6%)	2.95	21.57	44.67	25.70	5.10
1983 (n=21.657 =31.9%)	2.96	23.20	44.96	24.76	4.11

Oral exams:
(some students in
two counties)

1980 (n=1.263 =1.9%)	8.95	41.41	34.60	14.33	0.71
1981 (n=2.043 =3.0%)	6.22	48.21	33.72	11.50	0.34
1982 (n=1.330 =1.9%)	6.77	40.98	36.99	14.96	0.30
1983 (n=1.264 =1.9%)	6.65	40.03	40.11	12.82	0.40

The table makes it abundantly clear that the written exam dominates when seen in relation to the oral exam. Whereas between 30 and 40 per cent of the students sit for the written exam every year, only two to three per cent of the students sit for the oral. The individual student may thus for all practical purposes count on not having to sit for the oral. This backwash effect is serious when related to the official curriculum's emphasis on oral communication.

The distribution of marks for both achievement and written exam is surprisingly close to the normal distribution suggested in the regulations, particularly so for the written exam. In the achievement marks there is a weak tendency to avoid using the extreme marks. There is also a weak tendency to skew the marks toward the positive end of the scale, and the curve is generally somewhat flatter than officially suggested. Still there seems to be good reason to assume a relatively strong indirect effect of the standardized tests even if a normal distribution is no longer presupposed in the adapted goal-referenced evaluation. For the oral exam, however, the effect is not strong. Here about half of the students are found in the upper two bands.

On the basis of this material I would like to return to the question of teacher discontent with the goals-exam correspondence. Additional data are presented in Table 6. Here the teachers were asked if they think

that each of a number of areas is given the right emphasis in the present exam system.

Table 6. Relative emphasis on skills and knowledge in the present exam system. EFL teachers' views. Weighted and adjusted frequencies (n=191).

	alternatives					central tendency and dispersion	
	too much emphasis	-----			too little emphasis	Md	s
	5 %	4 %	3 %	2 %	1 %		
-written skills	19.6	34.4	34.2	11.2	00.6	3.617	.945
-oral skills	00.0	01.0	31.9	30.9	36.3	1.944	.850
-cultural knowledge	00.0	05.5	39.4	43.2	11.8	2.382	.767
-linguistic knowledge	00.7	04.2	70.4	20.6	04.1	2.868	.624

There is a major discrepancy between the evaluation of written and oral skills in the teacher material. Written skills are generally seen as having too much emphasis in the EFL exam, whereas oral skills are seen as having too little emphasis. In fact more than a third of the teachers feel that oral skills are given far too little emphasis. The median is as low as 1.994, indicating that this problem has considerable intensity (3). At this point it seems reasonable to suspect that at least one major cause of teacher discontent has been located.

The table also indicates that the cultural aspects of EFL are given too little weight. As to linguistic knowledge the neutral mid category is chosen by as many as 70.4 per cent of the teachers. In the Norwegian curriculum linguistic knowledge is deemphasized, and in recent exams it has not been tested in an explicit manner. It is interesting to note that so many teachers are satisfied with the exam in this respect. The standard deviation for this variable is lower than for any other variable we have analyzed so far. In other words there is both considerable satisfaction and agreement among the teachers on this question.

We have already to some extent considered the backwash effect of the exam. In a broader perspective different subtests may be evaluated as to their motivational effect on students. In the national survey the student respondents were asked about likes and dislikes as to selected test types.

Table 7. Written EFL test forms ranked according to grade 9 students' preferences. Weighted and adjusted frequencies (n=410).

	alternatives					central tendency and dispersion	
	like	-----	dislike				
	5 %	4 %	3 %	2 %	1 %	Md	s
comprehension questions	12.4	41.3	30.6	12.5	03.2	3.590	.971
fill-in-the-blank	15.3	33.8	40.9	08.4	01.6	3.477	.906
grammar questions	08.9	12.1	32.4	31.5	15.1	2.604	1.139
free essay	24.3	23.4	24.6	16.3	11.4	3.407	1.312
picture-based essay	08.9	23.7	36.6	18.7	12.1	3.025	1.126
narrative essay	09.3	20.2	27.6	24.6	18.7	2.756	1.231
expository essay	04.0	11.3	30.8	32.4	21.5	2.380	1.070
descriptive essay	04.4	17.9	21.7	35.6	20.4	2.331	1.333

From the students' point of view the essay part is the least popular. Discrete-point grammar questions are the only exception to this pattern. Both comprehension questions to a given text and subtests of the fill-in-the-blank type (including cloze) are generally well thought of.

Essays are not popular. However, the large dispersion both between and within essay types indicate considerable disagreement among the students. Free essays (where the individual student chooses her own topic) are generally well thought of. There are also groups of students who like other essay types.

In order to conclude the data section I would like to put the question of mixed-ability testing within the larger context of mixed-ability teaching. In a conference of testing we might easily forget that there are other and perhaps even more problematic areas to consider. The teacher questionnaires contained a section where the teachers were asked to evaluate their own teacher training with respect to a number of task areas relevant to their jobs. The results on a selected number of variables may shed some light on the question of evaluation versus other areas of teaching in a mixed-ability context.

Table 8. EFL teachers' evaluation of their own teacher training. Weighted and adjusted frequencies (n=191).

own training as to:	alternatives					central tendency and dispersion	
	very good	-----			very bad	Md	s
	5 %	4 %	3 %	2 %	1 %		
knowledge	38.8	38.6	16.8	05.8	00.0	4.210	.885
skills	29.4	41.2	24.9	04.4	00.0	4.001	.851
evaluation	05.9	09.9	36.0	22.9	25.2	2.553	1.147
differenti- ation	00.0	05.9	18.6	38.2	37.3	1.832	.892

It is evident that most teachers in the present sample are very well satisfied with both the skills and knowledge aspects of their own training. With the more pedagogical aspects of their training, however, the picture is different. Furthermore there is a clear difference between evaluation and differentiation in the data. Evaluation is a relatively mild problem after all in this context. Differentiation is a much more serious problem. This result is just one of several in the total survey material which suggests that from the teachers' point-of-view mixed-ability teaching is a far more serious problem in EFL than mixed-ability evaluation.

4 Concluding remarks

In the present paper I have tried to combine data from a national evaluation system with data from people working in the system. I think it is a fair conclusion after ten years of mixed-ability teaching and testing that such systems are quite possible to implement. Some aspects of these systems, as noted in this paper, actually seem to work well.

It also seems justified to conclude that when we choose to settle for a mixed-ability system we are not taking an easy way out as far as the organization of EFL is concerned. The system has its problems, particularly in practical teaching. These are going to be a continuing challenge for researchers and teacher trainers for a long time after the implementation of mixed-ability teaching and testing.

Haastrup (this seminar) has questioned the construct validity of the Danish oral exam in relation to the notion of communicative competence. This reservation is probably justified both for the oral and written exams in Norway. Here, the central problem seems to be

establishing both theoretically well-founded and precise definitions of central but vague terms like eg. 'coherence' (cf. footnote 2).

In Norway the communicative paradigm was not well established at the time when its curriculum was designed. It is to be expected that the 1985 revision of the teaching guide will change the basic non-communicative skills orientation into a more communicative one. This is hopefully going to lead to changes also in testing. In this connection, the material I have presented seems to call first for a reorientation in the direction of oral skills. It may also be hoped that the cultural aspects of foreign language teaching and learning will be emphasized more. Communicative competence is, after all, not simply a question of skill.

Notes

- 1 The statistics were produced by an old version of SPSS which does not compute percentiles.
- 2 It is significant to note that this term is not treated systematically in teacher training, teaching materials or actual teaching. At the University of Trondheim's Department of Applied Linguistics research is in progress to investigate students' written performance at the discourse-level. The research is based on the Trondheim Corpus of Applied Linguistics (Evensen 1982b), which consists of compositions written by the 2295 students who took part in the national survey. A central aim of the research is to carry out performance analyses to find correspondences between (non-) use of discourse-level features and holistic teacher evaluation in different grades. Preliminary results from exploratory studies are reported in (Evensen (in press) and in Evensen (forthc.)).
- 3 Considering both the well-known error of central tendency (cf. eg. Oppenheim 1966) and the fact that the measuring scale has only five steps, a median of 1.994 is very low indeed.

Bibliography

- Differensieringsfrågor i språkundervisningen.
Seminarierapport nr. 11. 1983. Helsinki: Nordiska språk- och informationscentret.
- Evaluerings i grunnskolen. 1980-1983. Grunnskolerådet.
Oslo: Grøndahl.

- Enkvist, N.E., ed. in press. Composition and coherence: a symposium. Publications of the Research Institute of the Abo Akademi Foundation. Abo: Abo Akademi.
- Evensen, L.S. 1982a. "Forskning for morgendagens fremmedspråkundervisning." In the report Lærere og skoleforskning. Oslo: NAVF.
- 1982b. "Data om språkundervisning". In Humanistiske Data, Vol. 9, No. 2, 4-11.
- 1984. "Difficulty in EFL: a national survey." In Ringbom & Rissanen (1984:349-358).
- in press. "Discourse-level interlanguage studies". To appear in Enkvist (in press).
- forthc. "Connecting L1 and FL in discourse-level performance analysis." Paper presented at the 20th International conference of contrastive linguistics, Poznan, Poland, Dec. 13-15, 1984.
- Frydenberg, R. 1981a. "Differensiering i språkundervisningen." In Undervisnings differensiering.... pp. 26-31.
- 1981b. "Målen och hur de uppnåtts samt utvecklingstrender i Norge. Undervisnings.... pp. 56-58.
- Frydenberg, R. 1981c. "Rollspel i undervisningen i främmande språk." In Undervisnings.... pp. 152-154.
- Håndbok for skolen, del III. Kirke- og undervisningsdepartementet. Oslo.
- Kunnskapsformidling i en skole for alle: innhold og nivå. OMI-rapport nr. 22. Oslo: Grunnskolerådet.
- Norsk Skole, 1982, 1984. Oslo: Kirke- og undervisningsdepartementet.
- Oller, J. 1979. Language tests at school. London: Longman.
- Oppenheim, A.N. 1966. Questionnaire design and attitude measurement. London: Heinemann.
- Ringbom, H. & Rissanen, M., eds. 1984. Proceedings from the Second Nordic conference for English studies. Publications of the Research Institute of the Abo Akademi Foundation, No. 92. Abo: Abo Akademi.
- Simensen, A. M. 1979. Evaluering i engelsk skriftlig ved avgangseksamen i 9-årig skole: en beskrivelse og en analyse. Report. Trondheim: University of Trondheim.
- 1983. "Intensjon og virkelighet i engelsk som fremmedspråk." In Kunnskaps-formidling.... pp.212-232.
- Stølen, L. 1984a. "Idéer og forslag til lærernes grunnutdanning og videreutdanning." In Differensieringsfrågor.... pp. 74-77.
- 1984b. "Resultat av forsøksvirksomhet innen undervisningens differensiering i Norge." In Differensieringsfrågor.... pp.78-79.
- Tangen, K. 1984. "Ressurser til grunnskolen i Norge." In Differensieringsfrågor.... pp. 80-107.
- Telhaug, A.O. 1970. Den ni-årige skolen og differensieringsproblemet. Oslo: Lærerstudentenes Forlag.

Undervisningsdifferentiering i främmande språk på grundskolans högstadium. Seminarierapport nr. 11. 1984. Helsinki: Nordiska språk- och informationscentret.

Utviklingstendenser i elevprestationer på ungdomstrinnet. OMI-rapport nr. 20. 1983. Oslo: Universitetsforlaget.

EFFECT OF MIXING ABILITY GROUPS ON ABILITY LEVELS ATTAINED

JOHN H.A.L. de Jong (Cito, Netherlands)

Summary

In the Netherlands the system of secondary education is highly differentiated. Possibly as a first step towards a comprehensive system two different types of secondary education were brought together in 1983. However, because of the width of the ability range, examinations at two different levels are offered at the end of the curriculum. The two levels were meant to be equivalent to those in the two types of secondary education of the former system. Data from samples of 300 to 1000 examinees on English and German foreign language listening comprehension tests were collected before and after the introduction of the new system. In an equating procedure based on Item Response Theory, using the Rasch Model, tests of different levels and from different years are interlinked by means of common items and common representative samples. The change in the educational system is found to have had a negative effect at both levels of achievement. These findings are corroborated by results on tests of other foreign language skills (reading comprehension) and other subjects (chemistry).

Introduction

Secondary education in the Netherlands is a highly differentiated system. Four different schooltypes aim at four different proficiency levels.

LBO provides lower vocational training and is attended by about 22% of the age cohort. LBO has a four year programme. Final examinations are taken as a rule at the age of 16. The foreign language programme takes about 300 hours and leads to a level comparable to the Threshold Level as defined by the Council of Europe (Van Ek, 1975).

MAVO is the lowest level of general education, meant as a basis for further training for clerical jobs and is attended by about 37% of the age cohort. MAVO has a four year programme. Final examinations are taken at the age of 16. The foreign language programme takes about 400 hours.

HAVO is an intermediate level of general education, meant to prepare for further studies leading to higher non-academic jobs and is attended by about 22% of the age cohort. Final examinations are at the age of 17. The foreign language programme takes about 500 hours and the final level is comparable to that reached by secondary education in most western countries.

VMO is the highest level of general education meant as a preparation for academic studies and is taken by about 18% of the age cohort. VMO is a six year programme. Final examinations are taken at the age of 18. The foreign language programme takes about 600 hours.

Within the MAVO schools, however, two subtypes have to be distinguished. Until 1979 pupils could choose either a three or a four year programme: MAVO-3 and MAVO-4. Final examinations for MAVO-3 were at a level below the regular (i.e.: MAVO-4) level. This MAVO-3 level was comparable to the LBO level and was taken by about 5% of all pupils attending MAVO. For several school subjects examination papers were identical for LBO and MAVO-3. In 1979 it was decided to bring all MAVO pupils together in a single school with the same four year curriculum for all. At the end of the curriculum pupils have the choice - for each subject separately - to take examinations at two different levels: MAVO-C or MAVO-D. These levels are meant to correspond to the MAVO-3 and MAVO-4 levels in the former system and several examination papers are once again identical for LBO and MAVO-C. The rationale behind the change of the system for MAVO is to offer all students better chances. A MAVO-3 certificate was not greatly appreciated. It was therefore felt to be better for MAVO students to be able to postpone the definitive choice for a particular level; to offer all MAVO students the opportunity to obtain the higher level MAVO certificate. If this aim should prove to be too high for some, then the possibility of taking exams in a number of subjects at the lower level would give them a chance to obtain a certificate with the prestige of the higher level MAVO for at least the remainder of the subjects.

In 1983 the first cohort of MAVO pupils in the new system presented themselves for final examinations. The present study was undertaken to reveal the effect of the changes in the MAVO schools on the achievement level of MAVO pupils: did the postponement of level determination and the lengthening of the curriculum for the lower level lead to better results and to higher certification?

Method

The method used is based on Item Response Theory using the Rasch Model (Rasch, 1960). In the Rasch Model one item parameter and one person parameter determine the probability that a given person answers a given item correctly. The item parameter in the Rasch Model is called the difficulty parameter and the person parameter is called the ability parameter. However, since for the calibration of each test an arbitrary origin of measurement is selected the numerical values of the parameters of different tests cannot be compared directly. In order to make them comparable a common origin has to be determined for the different tests. This procedure, known as test equating, can be followed if:

- the tests contain a number of common items;
- the groups taking the tests contain common persons;

- the groups taking the tests are representative samples from the same population.

If one of these conditions is met, and if the tests measure the same ability, and the tests are not too far apart in difficulty and reliability (Petersen et al., 1982) the item parameters of the tests and the ability parameters of the persons can be equated, i.e. the parameters can be put on a single scale.

In 1982 (the last year of MAVO examinations in the former system) random samples of 300 were taken from the LBO, MAVO-3 and MAVO-4 populations. Because the LBO and MAVO-3 levels are supposed to be identical LBO and MAVO-3 students take the same foreign language listening comprehension test. The students in each sample were asked to take one third of the test of the other level besides taking the test at their own level: thus LBO and MAVO-3 students took a part of the MAVO-4 test and MAVO-4 students took a part of the LBO/MAVO-3 test. In 1983 and 1984 similar procedures were followed with samples of 300 to 1000 from the LBO, MAVO-C and MAVO-D populations. This design would permit equating both the tests and the populations concerned by means of a common item link. Random samples of 1000 LBO pupils from the years 1982, 1983 and 1984 were taken to form representative samples of the same population, thus permitting to equate the scales of the subsequent years.

The tests used in this study are the regular Cito tests of foreign language listening comprehension. The tests were originally developed in a research project at the University of Utrecht (Groot, 1975). Because examinations are public, new tests have to be developed each year. Cito tests of foreign language listening comprehension have been constructed at Cito and used in final examinations in schools for over ten years. The objective of the tests, as defined by Groot (1975), is to evaluate foreign language learners' ability to understand the foreign language, spoken spontaneously by educated native speakers at normal conversational speed. Extremely informal elements as well as lexical and syntactical elements that are incomprehensible to less educated native speakers and topics requiring specific knowledge are excluded from the tests. The language material consists of three or four interviews with native speakers of various occupations and professions. The item format is a multiple choice question with three options. The language material is divided in samples of forty to fifty seconds; the correct answer is a one phrase summary of the global contents of the sample. Each sample is followed by a pause of twenty seconds on the tape which allows the students to read the item in their test book and tick their answer on an answer sheet. A second item format was introduced (De Jong, 1983; De Jong, 1984): modified cloze items with two options. In each sample of twenty to thirty seconds on the tape a word, or group of words is deleted and replaced by an electronic signal. Words to be deleted from the original text are chosen for their semantic relevance in the context. The language material consists of discussions, conversations, news items and other radio programmes. The students are to decide in a pause of seven seconds

following the language samples on tape, which of the two options in their test book can be used to restore the original text. Depending on level and item format total test length may vary from forty to fifty items. Administration time varies between fifty and sixty minutes. Test reliability (KR20) ranges from .75 to .85.

The test analyses according to the Rasch Model were performed with the computer programme CALFIT (Wright and Mead, 1975). Tests were equated following to the procedure described by Wright and Stone (1979).

Results and discussion

Figure 1 presents the results from the equating procedures on the German foreign language listening comprehension tests for MAVO in the years 1982, 1983 and 1984. The three horizontal lines in figure 1a represent the ability variables in the subsequent years, but should in fact be regarded as a single line as the tests have been calibrated on a single scale. Above the lines the mean ability at the higher MAVO level is indicated by triangles. The shaded areas represent the distribution of ability from two standard deviations below the mean to two standard deviations above the mean, thus covering about 96% of the population. Arrows indicate the cut-off point. Below each line the same is shown for the lower MAVO level. The dot on each line represents the LBO mean ability in the subsequent years, the dots are connected by a dotted vertical line. The figure shows that there is a difference of about one logit (the unity of measurement on the ability scale) between the two MAVO levels. This difference remains fairly constant during the three years. The mean ability of both MAVO levels shifts to the left, that is: to the lower end of the ability scale in 1983. Although the mean ability at both levels slightly moves to the right again in 1984 it does not reach the original level of 1982.

The number of students at the two MAVO levels is not equal and varies from year to year. A larger proportion of the total MAVO population has taken examinations at the lower MAVO level from 1983 on. This has an effect on the mean ability of the total MAVO population. The mean ability and the distribution of ability in the total MAVO population is pictured in Figure 1b. Figure 1b is drawn at the same scale as Figure 1a. The arrows indicate the mean ability in the subsequent years; the shaded area represents the distribution of ability of 96% of the total MAVO population (both levels). Apart from the shift to the lower end of the ability scale a smaller spread in ability can be observed.

Figures 2a and 2b picture the same kind of data for English listening comprehension. Figure 2a shows a comparable shift to the lower end of the ability scale as was found for German. For English, however, the mean ability of the lower MAVO level, which was well above the mean LBO level in 1982, is below LBO in 1983. Again, as for German, a slight shift to the right in 1984 does not result in

complete recovery of the original level. The combination of the two MAVO levels in Figure 2b shows a similar narrowing of the spread in ability of the total MAVO population as was found for German.

An observed shift in mean ability of a population of about 90,000 students cannot be interpreted as a standard error of the mean. Furthermore, it is highly improbable that such a standard error of the mean would cause deviation in the same direction for two different years and for two different languages. To check on sampling errors reference was made to the results of the populations concerned on reading comprehension tests. Results on the reading comprehension tests are collected each year from the total population: 40,000 LBO students and 90,000 MAVO students. Figure 3 pictures the relative position of the lower level MAVO students and the LBO students from 1978 to 1984 in standard deviations of test scores. The mean LBO score is set at zero, the horizontal axis represents time and the vertical axis standard deviations of test scores on the reading comprehension tests. The mean score of the lower level MAVO students for English reading comprehension was traditionally well above the mean score of LBO students. This situation changes in 1983: the lower level MAVO students drop to a level that is more than half a standard deviation below their former level. For German a similar drop can be observed, in this case enlarging the original distance between LBO and lower level MAVO.

Another possible influence on the mean ability of the MAVO and LBO students could be a change in the groups involved. If quite a large proportion of the age cohort has chosen for a different schooltype since the change of the system, a shift in mean ability of the groups is to be expected. Figure 4 pictures the number of examinees taking English (ca. 97% of the total population) in the different schooltypes from 1972 to 1984. Looking at the period from 1972 to 1982 a constant growth, parallel to the postwar increase in population can be observed. The two MAVO groups are an exception in that the lower level MAVO group decreases from 1980 on going with a more rapid growth of the higher level MAVO. Immediately after the change of the MAVO system a dramatic change in the two MAVO populations can be observed: more than three times as many MAVO students opt for the lower level MAVO. There is no indication of a shift in population between LBO and MAVO. The larger number of lower level MAVO students is apparently recruited from the group that formerly opted for the higher level MAVO and would therefore be expected to raise the level of achievement in the lower level group. Furthermore, hypothesizing that it will not have been the highest achievers in the higher level group that deserted this level for the lower, one could equally have expected a rise in mean ability for the higher level MAVO. The observed change in numbers of the MAVO groups therefore, cannot account for the downward shift in achievement. On the contrary, this change would lead one to expect a shift in the opposite direction. Bakker (1984) reports that the hypothesis of a rise in mean ability of the higher MAVO group in chemistry, due to the changed distribution of MAVO students over the two MAVO levels

after the change of the system, has to be rejected.

Apart from the effect on the achievement level of MAVO students the change in the MAVO system has had its effect on their level of certification. Figure 5 pictures this effect. The proportion of MAVO students taking a subject at the higher MAVO level has gone down for all subjects. Figure 5a shows the exact proportions in 1982 and 1983 for English. For the foreign languages the proportion of ca. .05 of the total MAVO population that took the lower MAVO level has gone up to a proportion of .15 to .18. For subjects such as mathematics and chemistry, traditionally regarded as difficult, the original .05 may go up to proportions of .20 to .40.

The advocates of the change in the MAVO system present the statistics of Figure 5b as proof of a positive effect. Before the change of the system 15% to 20% of the total MAVO population finished their last year without any certificate at all and had to sit the last year again to obtain a certificate, which meant the loss of an entire year for these students. In the new system, only about 7% or 8% fail to obtain a certificate. However, one whole year has been added to the curriculum for about 5% of the students: they have had their "loss of a year".

To make a fair comparison between the certification of MAVO students before and after the change of the system, the level of certification also has to be taken into account. Figure 5c shows that the proportion of MAVO students taking all subjects at the lower level has remained more or less the same in 1982 and 1983. The number of failures amongst this group has increased dramatically: from about 20% in 1982 to almost 50% in 1983. The proportion of candidates taking MAVO exams in all subjects at the higher level in 1983 is only half of that of the years before. Granted, the students taking all subjects at the higher level do so with success in 1983 a larger proportion passes the exams, but it is much smaller group than before. A smaller proportion of MAVO students will in effect enjoy the greater social appreciation associated with the higher level certificate. The proportion of MAVO students passing at the higher level MAVO for all subjects has gone down from about .80 in 1982 to about .53. Between lower level MAVO and higher level MAVO intermediate levels have been created; students can now have mixed level certificates. These mixed certificates can contain one subject at the lower level and five subjects at the higher level; or five subjects at the lower level and one subject at the higher level or any other combination with a total of six subjects. How exactly the mixed level certificates will be appreciated cannot be foreseen, but a certificate with all subjects taken at the higher level will no doubt receive more appreciation.

After only two years it is too early to pronounce a final judgement on the new MAVO system. However, the downward shift in the mean ability of the total MAVO population for listening comprehension of German and English, the relatively lower scores for the lower level MAVO on reading comprehension tests of German and English, the absence of a rise in mean ability in chemistry in spite of a more

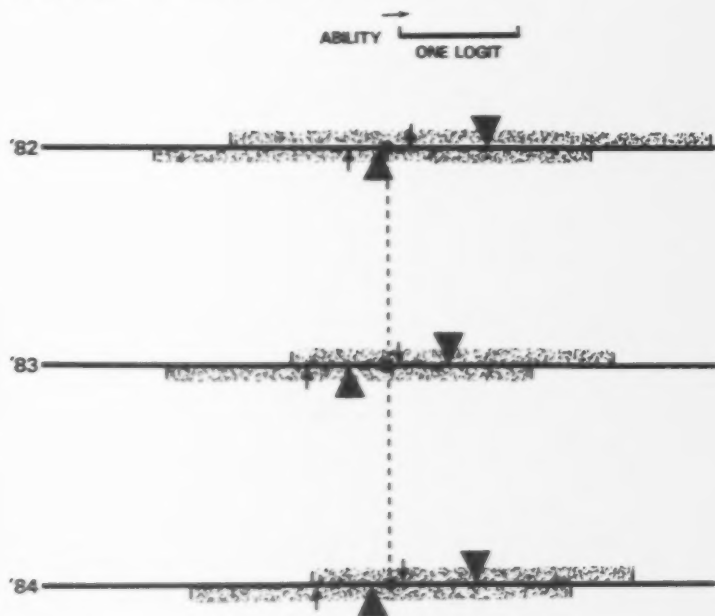
select group, the smaller proportion of MAVO students obtaining a certificate with all subjects taken at the higher level, all these observations reveal a similar tendency: mixing ability groups has a negative effect on the final ability attained. Various explanations for this negative effect can be offered. Teachers may tend to adapt their level of teaching to the mean level of the students in the group. Students may be less motivated to do their utmost when they know failure at the higher level does not mean total failure: they can always try their success at a lower level. Though more data on the results of the new system are needed, it is impossible to welcome the present results and extreme caution has to be taken in further experiments towards a comprehensive system.

References

- Bakker, S. 1984. MAVO-niveau op drift. In: Weekblad voor leraren van het N.G.L. 17, 8, 298-299.
- De Jong, J.H.A.L. 1983. Focusing in on a latent trait; An attempt at construct validation by means of the Rasch model. In: J. van Weeren (ed.), Practice and problems in language testing 5: Non-classical test theory. Final examinations in secondary schools. National Institute for Educational Measurement, Cito, Arnhem.
- De Jong, J.H.A.L. 1984. Testing foreign language listening comprehension. In: Language Testing 1,1,97-100.
- Groot, P.J.M. 1975. Testing communicative competence in listening comprehension. In: R.L. Jones and B. Spolsky (eds.), Testing language proficiency. Centre for Applied Linguistics, Arlington, Virginia.
- Petersen, N.S. Marco, G.L. and Stewart, E.E. 1982. A test of the adequacy of linear score equating models. In: P.W. Holland and D.B. Rubin (eds.), Test equating. Academic Press, New York.
- Rasch, G. 1960. Probabilistic models for some intelligence and attainment tests. Danmarks Paedagogiske Institut, Copenhagen.
- Van Ek, J.A. 1975. The threshold level. With an appendix by L.J. Alexander. Council of Europe, Strasbourg.
- Wright, B.D. and Mead, R.J. 1975. CALFIT: Research Memorandum nr. 18. Department of Education, University of Chicago, Chicago.
- Wright, B.D. and Stone H.H. 1979. Best test design: Rasch measurement. Mesa Press, Chicago.

Figure 1 German listening comprehension 1982-1984 :
Mean ability and distribution of ability of MAVO candidates

1a At two MAVO levels separately



1b Both MAVO levels combined

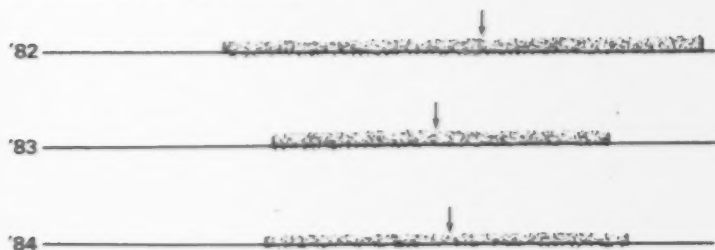
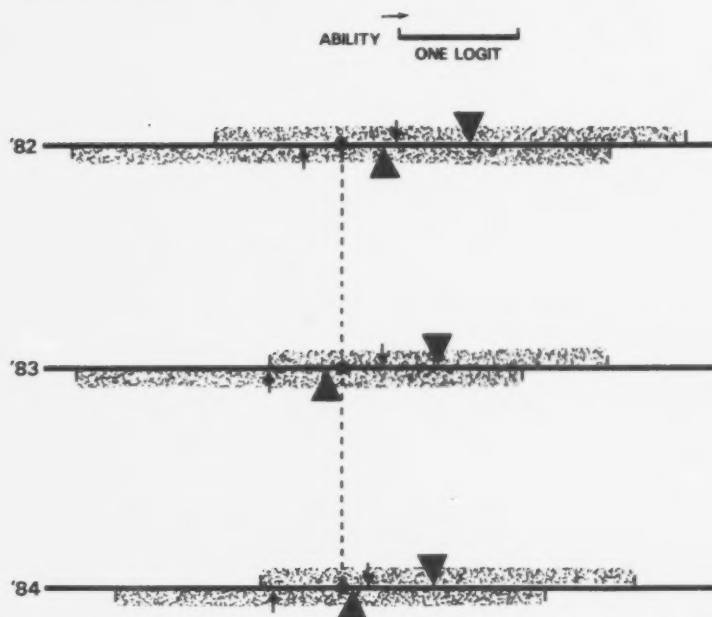


Figure 2 English listening comprehension 1982-1984 :
Mean ability and distribution of ability of MAVO candidates

2a At two MAVO levels separately



2b Both MAVO levels combined



Figure 3 Reading comprehension 1978-1984 :
Relative achievement level of LBO and MAVO candidates

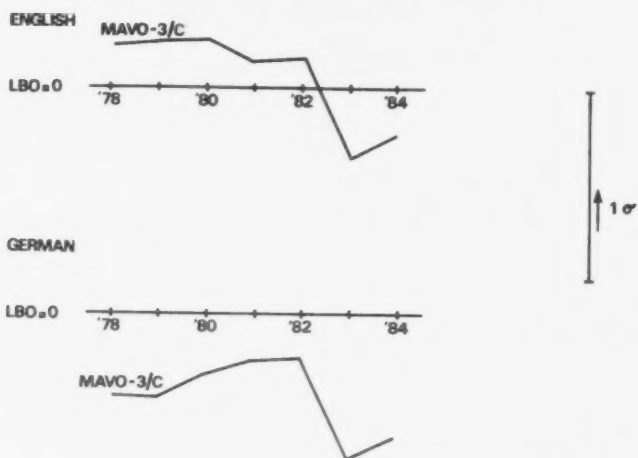


Figure 4 Number of examinees taking English: per schooltype 1976-1984

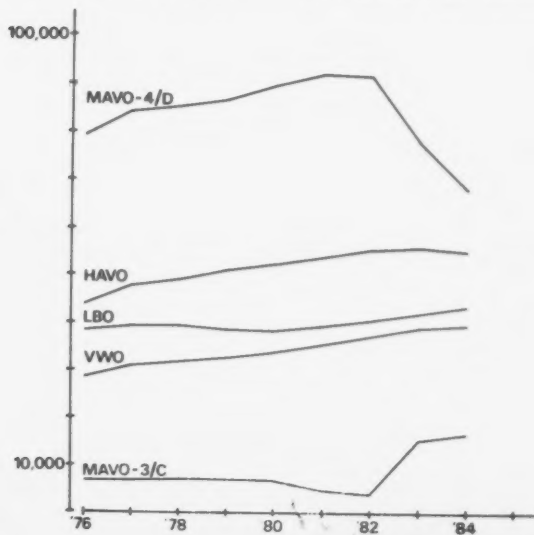
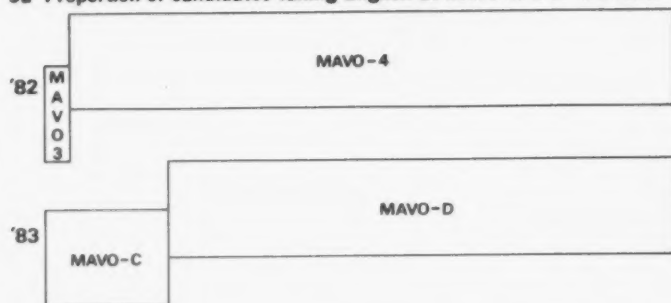
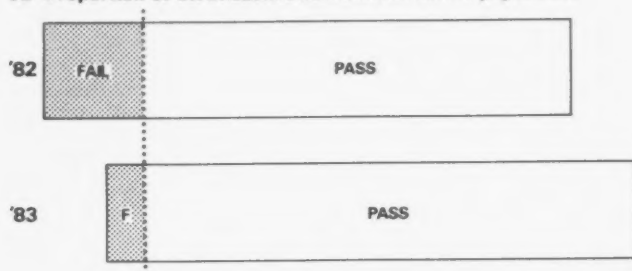


Figure 5 MAVO examination levels attained in 1982 and 1983

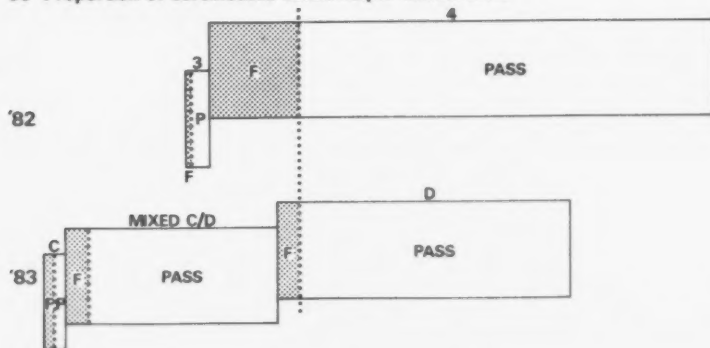
5a Proportion of candidates taking English at MAVO 3/C or 4/D level



5b Proportion of certificates attained:total MAVO population



5c Proportion of certificates attained:per MAVO level



TESTING COMMUNICATION SKILLS IN MIXED-ABILITY GROUPS

Viljo Kohonen (University of Tampere)

Abstract

This paper will first discuss some aspects of education, evaluation and communication with regard to the instructional goals of the Finnish comprehensive school. Problems involved in designing summative tests for mixed-ability groups will be discussed next, and an example will be given of a test of English administered to 7th-graders (N=104). It is suggested that communication is a "graded" property that is possible at various levels of accuracy. The essential criterion in scoring should be comprehension. If the development of communicative skills can be seen as a continuum ranging from "zero communication" to virtually bilingual competence, learners can and should be encouraged to work towards the quality of skills that is within their reach.

1. Some views of education and evaluation

1.1. Evaluation to improve learning

It is occasionally salutary to ask oneself the creative question, why test at all? Thought-provoking ideas have been offered by Bloom (1981), who makes a distinction between two major functions of education: selection and developing learner talent. Bloom points out that education has traditionally had a selective function, whereby the basic task of education was the "identification of the few who were to be permitted to enter and complete the secondary school academic program and then be admitted to higher education" (p.2). In this view, a central task of education is to classify learners for various "streams" of learning, under the assumption that learners are different and the school must select those that are capable of higher education. This guiding function is, of course, still one task of the school. But Bloom argues that it has become less important in developed societies in which a major proportion of the age group completes secondary education and there is an increasing demand for education for all citizens in a complex society. In this situation, education should primarily serve the development of the individual. Education should develop those characteristics in all students which will enable them to live effectively in a changing society. Thus the school should devote major resources to increasing the effectiveness of individuals rather than predicting and selecting talent:

Education must be increasingly concerned about the fullest development of all children and youth, and it will be the responsibility of the schools to seek learning conditions which will enable each individual to reach the highest level of learning possible for him or her. (p.3)

Bloom further points out that the rapid change in modern society requires education to continue throughout life. Education must thus be seen as a lifelong process that never fully ceases (cf. Faure et al., 1972). Now,

these views of education are closely connected with three philosophies in evaluation: the selective, criterion-referenced and "learner-supportive" evaluation (cf. Warries 1982). The central question is what the learner's score is compared to: whether the scores of his peers, a pre-defined mastery level, or his previous learning.

The selective philosophy is, of course, well-known to all teachers and is deeply rooted in differential psychology and the statistical trend for the "normal" distribution of human abilities in large groups. It is connected with the selective function of the school. What is considered excellent, moderate or poor performance is determined by the empirical distribution of the scores. By definition, an "excellent" performance by some learners implies less excellent performance by some others. One can ask with good reason what are the affective consequences of the comparisons for the less able learners (cf. Bloom 1971a), and how necessary such comparisons really are in education. Information about rank order is obviously needed for admission to vocational institutions, educational programs or universities, i.e., in instances of restricted intake to further education. But such information is not needed within the institution itself.

In the criterion-referenced philosophy, the learner's performance is compared with a required level of mastery, based on the objectives of instruction. Instead of competing with his peers, the learner competes with the objectives of the learning task (cf. Takala 1985; Carroll 1971; Bloom 1971b, 1976; van der Linden 1982). This philosophy represents a vast improvement in educational thinking. Learning is seen as an individual effort of the learner towards the objectives. There are, however, some shortcomings: the definition of the behavioral domain is problematic, as is the definition of the required level of mastery (cf. Takala 1985). Further, in school conditions with fixed numbers of teaching periods available, it seems unavoidable that a number of learners will not reach the mastery level, while others will go beyond the common objectives. A problem for fast learners is, however, whether they have enough incentive to work for the "extended" objectives, after reaching the common objectives. Is there a danger perhaps that they may not work as hard as they could?

This is where the third, learner-supportive philosophy of evaluation would seem to offer some promising possibilities. In this approach, the learner "competes" with himself, to reach his personal objectives and augment his learning in comparison with his previous level of learning. This is essentially what Bloom seems to suggest in his second view of education: enabling each individual to reach the highest level of learning possible for him. This thinking leads to individualized learning whereby learners take on "learning contracts" that are reasonable within their total learning situation. In a sense, any learner that has efficiently utilized his potential is a "good" learner. This suggests bigger contracts for fast learners and smaller ones for slow learners. Such contracts should not, however, be forced on the learners, but should be based on their own responsibility and thus gradually lead to an attitudinal growth towards their being autonomous learners.

It is perhaps impossible, and even undesirable, to define exactly

what the "efficient" use of the potential means. Rather than viewing it as a "closed" phenomenon, it should be seen as an unfolding potential which depends on genetic, environmental and individual factors. It refers to the learner's total learning situation and includes ability factors, learning opportunity, home and school support, and the individual's needs and motivation (cf. Strevens 1980). We should beware of labelling our learners too hastily, as teacher expectations may involve well-known "self-fulfilling prophecies" (cf. Rosenthal and Jacobson 1968; Elashof and Snow 1971; Rosenthal and Rubin 1978; Babad et al. 1982). We should rather aim towards an open-ended, basically supportive view of our learners and encourage them to go as far as they wish and are able to go in their learning, within the possibilities allowed by their total learning situations.

1.2. On goals of education and foreign language instruction in Finland

In the new Comprehensive School Law (1983), the educational goals are defined in terms of a number of desirable learner characteristics to be attained by the end of the obligatory school. Specifically, the school should aim to educate pupils so that they are

- well-balanced
- responsible
- autonomous
- creative
- cooperative
- peace-loving

and physically healthy individuals and members of society. In addition, the school must provide education in morals and good manners and impart knowledge and skills that are necessary in their lives (§ 2). Language teaching aims are specified in the recent syllabuses (1982), whereby the general aims of foreign language teaching are defined as the "useful basic skills needed in everyday situations", a "positive attitude to language study" and a "continuous maintenance and development of the skills after school". Language teaching should also impart knowledge about the cultural background, increase international understanding, and further the pupil's ethic, aesthetic and socio-emotional growth. Language teaching is thus seen as language education, aimed at increasing the learner's skills of self-expression. Every language teacher is also a language educator. The syllabus contents include common objectives for all, and extra materials are offered for fast learners.

1.3. Learner communication as a "graded" property

Even though all learners work on the same common objectives, it is obvious that not all of them will reach the same level of communicative skills within the same amount of instructional time - there will always be learners for whom languages are easy and those for whom they are difficult. Thus they will end up with various levels of proficiency at the end of school. It is particularly the abstract rule system that poses difficulties for slow learners, while fast learners are still able to acquire both accurate and fluent language skills. Due to the historical tradition, we have been too much concerned with accuracy as the language teaching objective; we all know, of course, that communication is possible even with erroneous code. Errors will variously hamper communication and may

even cause it to break down in some more serious cases, but the essential criterion for communicative success must be comprehension and not grammatical accuracy.

Communication is a "graded" property which is possible at various levels of grammatical and lexical accuracy, ranging from mere non-verbal communication to fluent verbal and non-verbal communication. In fact, it is deficient lexis that hampers comprehension more than inaccuracies in pronunciation or grammar (cf. Littlewood 1984; Dulay et al. 1982). Accuracy should obviously be the goal to work towards, but it is not an end in itself. Equally important is that learners should have something to say and the courage to use the language, even at the risk of making mistakes. Accuracy is nothing more than a tool for efficient communication.

If accuracy may not be reached by all learners, everybody should be able to communicate something in the foreign language nonetheless. Thus some acceptable level of communication should be within everybody's reach. Now, communicative thinking would seem to offer a possibility for reconciling variation in learner ability and instructional objectives. We might consider how far it is advisable to try and push the abstract system with slow learners, knowing that communication is not a simple matter of "success or failure" in interactive face-to-face situations. As Henry Widdowson (1979) has repeatedly pointed out, it is a matter of negotiation about meanings. Thus, if I do not understand what is said to me I can always ask for clarification, just as my interlocutor can do when I cannot make my meaning clear. Successful communication requires courage, perseverance and communicative self-confidence. And such attitudinal objectives should also be included in our teaching and testing syllabuses.

Foreign language teaching should naturally provide normative feedback to learners, since they need to know how they are progressing in their learning tasks. But language testing need not be geared strictly to requirements of accuracy for all learners. Thus slow learners can - and should - be given some credit for any comprehensible language and effort. But by the same token fast learners should be required to attain high standards of accuracy, and they should be encouraged to work to this end. This means, then, accepting different learning outcomes for different learners, depending on their ability (and motivation) to learn languages.

2. Testing language skills in mixed-ability groups

2.1. Some guidelines for test design and scoring¹

With the advent of the new Finnish school laws (from August 1985 onwards), mixed-ability teaching will be a reality in foreign languages and mathematics in the upper comprehensive school (grades 7-9) as well. This means that all learners will be offered the same common objectives within the same amount of time, and their skills will be tested using the same tests for all. The change has been preceded by a number of teaching experiments in various schools, in order to develop pedagogical solutions, testing techniques and learning materials. In our teaching experiments in Tampere, we have found that the following principles can serve as useful guidelines for designing and scoring tests for mixed-ability groups.

(1) What to measure? This is, of course, the basic question of the validity of the test. The central question is, what we mean by "knowing a language" and how we operationalize our views in concrete tests. Summative tests will inevitably indicate what we regard in our syllabuses as "worth learning". In communicative thinking this must be essentially the ability and willingness to put into a communicatively meaningful use whatever amount of language the learner has acquired. Tests must therefore be designed so that they give the learner opportunities to show what he can do with his language. This brings in the requirement of authenticity or, in any case, an attempt to simulate real-life tasks of language use, which is the goal for which we must prepare our learners. Tests must also be fair. That is, they should not contain any unnecessary gimmicks or tricks aimed at "catching" weaknesses in the learner's skills. While testing is always artificial to some extent, we should attempt to reduce artificiality as far as it is possible within the practical constraints of the classroom situation and the available resources.

(2) What kinds of tests to use? This question is a corollary to the above question of validity. Communicative skills can be measured by various kinds of tests, calling into play different aspects of communicative ability. We should therefore aim at using various types of tests, in order to get a many-sided coverage of the skills involved. Any single test type (or a limited number of types) will inevitably give a narrow measure of the skills. We should not lose sight of the communicative relevance of the tests. This is also important because of the well-known backwash effect of tests on teaching: our testing procedures should encourage communicatively oriented classroom work. Analytic tests could be used as formative tests of the learners' knowledge of the discrete points of grammatical structures. Integrative tests, on the other hand, would be better suited as summative tests, enabling learners to show what they can do with their language skills. This would provide a communicatively balanced basis for giving grades in the school reports.

(3) At what level of mastery? Following the familiar Bloomian taxonomy (cf. also Takala 1985), language processing can be thought of in terms of three levels:

- recognition of language items, understanding of meaning units
- mechanical skills, limited recall, guided production
- creative, autonomous, personal use of language in communication

For slow learners, simple recognition is already quite demanding, and they seem to need cues to guide their productive use of the language. Recognition of discrete structural forms and meanings by answering multiple-choice items is, of course, a relevant part of language competence, but this is rather a limited view of the variety of the language skills needed in real-life contexts. Such contexts will typically require open-ended, independent interpretations of meanings. Besides, if we think of language processing as a creative reconstruction of meanings, we also have to provide open items to measure such skills. Perhaps comprehension can be measured more directly by open tasks, since a learner who is able to answer open questions will usually pass multiple-choice items as well, while the converse is not necessarily true (for the well-known distinction between recognition and recall, cf. Anderson 1980). In this sense open questions are more economical, as they measure language processing at a deeper and more demanding level.

(4) How to score learner productions? Scoring problems are bound to arise in mixed-ability groups where learner responses will scatter over a wide range in terms of both quality and quantity of the language. To cope with the variation, some scale is necessary in scoring; mere 0/1 will not do justice to all. Besides, communicative thinking also suggests that language use is not a simple "wrong/right" matter; the same contents can usually be expressed in a number of possible ways, and what is involved is rather a scale of acceptability and communicative efficiency. The basic criterion for the application of the scale must be comprehensibility: has the learner understood the gist of the text? or would a speaker of the target language understand the intended meaning? If the answer is positive - despite incorrect spelling and bad grammatical mistakes - the learner has managed to interpret or process the message successfully and deserves some credit for it. Obviously, a zero will be given for a blank answer as well as for an answer that is clearly not interpretable. Thus, on a scale 0 - 3, for example, 3 points will be given for a "perfect" answer, 1 point for a comprehensible attempt, and 2 points for a variety of combinations of partly incorrect language and deficient message contents. There is no reason why slow learners could not answer in their mother tongue in listening and reading comprehension tasks if they have difficulties in writing. What is measured is their ability to understand target texts, and this can equally well be controlled by answers in the mother tongue. There is no reason to be dogmatic about the use of the mother tongue. More important is the communicative quality and meaningfulness of the tasks, and the learner's motivation for and involvement in completing them.

(5) How to transform the scores into grades? Communicative thinking leads to a criterion-referenced orientation, with communicative efficiency and the degree of attainment of the objectives as the criteria. The specification of the mastery level is a difficult task, and the application of the criteria to the evaluation of the learner language is always a matter of subjective interpretation. This entails problems of both intra-rater and inter-rater reliability. But such problems should not be exaggerated. It is a well-known problem that validity and reliability are difficult to maximize in the same test. While multiple-choice tests eliminate problems of inter-rater reliability, their validity is limited as a measure of the communicative use of the foreign language. Of the two important statistical properties, validity must be given the first priority in communicative testing. On the other hand, any communicative test must also meet sufficient requirements of reliability. Reliability can be increased by specifying the criteria and ensuring their consistent application by different raters.

In school contexts, evaluation usually involves a conversion of the scores into some scale of grades in the school reports. As noted above (Section 1.1.), relative grading should not be considered so "necessary" while the students are in school - society and higher institutions will need such information only in school-leaving reports. More important than learner comparisons is the tenet that evaluation should be geared to improving student learning. It seems very difficult to convert the learner-supportive evaluation philosophy into any commensurable scale of numerical grades that could be applied consistently to all in mixed-ability groups, but this might function as a framework for self-assessment by the learner himself. One possibility for converting the scores into school grades within a somewhat loose criterion-referenced orientation would be

to describe the pass grades (in the Finnish system, grades 5 to 10) at three broad levels of performance:

- (1) level 1: "pass level" (grades 5 and 6)
- (2) level 2: "common objectives" (grades 7 and 8)
- (3) level 3: "extended objectives" (grades 9 and 10)

For the lower grades, accuracy standards are less important as long as communication is comprehensible, while accuracy is an important criterion for the two highest grades (9 and 10). Fast learners will thereby be encouraged to work both for accuracy and fluency. They can take on more demanding learning contracts by aiming towards qualitatively higher standards of language use. In addition to the written tests, the grades are also affected by other factors, such as oral skills (tested mainly through continuous assessment), contribution to classroom work, effort to learn, and diligence in doing homework. Language testing is thus seen as part of the school's wider educational task and the grades are also based to some extent on the educational goals discussed above (Section 1.2.).

2.2. Some empirical findings of a written English test (7th grade)

One aim of the Tampere experiments has been to develop possible testing techniques for mixed-ability groups. An attempt towards this aim is the test in English (given in Appendix 1), administered in Autumn 1983 in grade 7 (i.e., after some 300 lessons' instruction in English, learner age 14 years, N=104). The test consisted of the following parts:

- (1) contextualized vocabulary test (20 items)
- (2) contextualized past tense recognition and production (16 items)
- (3) cued interview - ask five questions (5 items)
- (4) cued dialogue completion (5 moves to complete)
- (5) telling about past events (6 scorable points)
- (6) reading comprehension test (5 open questions in Finnish)

The total number of items was 57, with the maximum score of 90 points. The statistical results of the test are given in Appendix 2.

In the vocabulary test, the first eight items were aided recognition in context: the learners were given the Finnish word in brackets, with the English counterparts in the box on the exam sheet. They had to be able to identify the right word out of a large selection. In items 9-15, the Finnish words were still given, but the English counterparts had to be recalled. Items 16-20 were a cloze test, with acceptable-word scoring (0/1). The statistical analysis showed that the test functioned in a satisfactory way: the items were progressively more and more difficult (solution percentages from some 80 to some 30 per cent), and the reliability index was high (.94, internal consistency, Cronbach's alpha coefficient).

Test 2 (past tense formation) was easy (average solution percentage 76.4%) but functioned quite well, becoming more difficult towards the end (reliability index .90). Test 3 (cued interview) was scored on a 0-2 scale, to give some credit for partially comprehensible answers. It proved similarly quite easy (71.1%) and functioned well (reliability index .88). Tests 4 (dialogue completion) and 5 (telling about past events) were scored partly on a scale ranging from 0 to 3. They were more difficult (66.6% and 55.8%), but functioned well, particularly test 5, where the reliability index (.93) was good considering that the test consisted only of

6 items. It seems that open questions are an efficient measure of language skills. Even though they are more laborious to score, the amount of work is compensated to some extent by the fact that the test need not be long in order to be reliable. The reading comprehension test had two items (4 and 2) which were too easy and actually measured pragmatic knowledge rather than comprehension of the text; its reliability remained poor (.65). The reliability index of the total test battery was very high (.98), indicating consistent discrimination. Thus it can be said that the test as a whole functioned well statistically.

The frequency distribution of the scores shows that the test was relatively easy. The shape of the diagram is, in fact, a typical mastery learning curve which is positively skewed. Thus a majority of the pupils obtained quite high scores: the average score was 60.5 (with $s = 23.8$), which is 67.3 per cent of the maximum total score (90 points). The conversion of the scores into the grades was done so that the cut-off points for the grades were certain proportions of the learner score out of the maximum total score. The cut-off points for the four broad levels of the grades were as follows:

score:	level	grade
below 24 points	Fail	(4)
24 - 56	Pass	(5 - 6)
57 - 78	Good	(7 - 8)
79 - 90	Very good	(9 - 10)

It will be seen, then, that the scale was more lenient at the lower end of the score distribution, while very good grades required accurate language use as well. The cut-off points were arbitrary and based on traditionally used proportions of the learner scores out of the total score. They were thus not based on mastery descriptions, while attempts were made to ensure that the common objectives were reached by those learners who got a "good" grade (above 7), i.e. beyond 57/90 points (thereby getting a minimum of 63% out of the total score).

These results indicate, then, that the present test is one possible way of measuring written language skills in mixed-ability groups. But it is just one possibility that needs to be developed further. In retrospect, it is still too much oriented towards testing grammatical structures (in this case, the past tense). It would seem more advisable to test the mastery of analytic grammatical structures in formative tests and thereby shift the emphasis in summative tests towards evaluating the use of global skills. As noted above, this would give a communicatively balanced basis for the grades in the reports. To make the tests communicatively more meaningful for the learners, more opportunities should also be provided for a personal and creative use of the language.

Another shortcoming is the lack of a systematic evaluation of the spoken skills. Any written test is bound to give only indirect evidence of the learner's abilities for face-to-face interaction. As this would involve arranging interview tests, which are quite laborious in large learner groups, it has not been possible to do it so far. Thus the testing of spoken skills has remained more informal, on the basis of teacher observation during the lessons. Perhaps it will be possible to make such observations more systematic and thereby more reliable by developing criteria for continuous assessment.

3. Discussion

When a foreign language is taught to the whole of the age group learner performance will be graded in terms of both the quality and quantity of the language produced, as it is in the mother tongue communicative skills as well. On the other hand, communication is also a "graded" property in the sense that it is possible at a variety of accuracy levels. Errors will variously hamper communication, but language is so redundant that it "tolerates" a fair amount of mutilation before communication really breaks down. Besides, in normal interactive situations, it is usually possible for the interlocutor to check his comprehension by asking. Language use is thus a matter of negotiation about meanings. While language teaching should naturally be normative and thereby have accuracy as an aim, accuracy should not be seen as an end in itself, but rather as a tool for efficient communication.

These ideas about learner variation and communication could lead to a somewhat individualized approach whereby different learners take on, as it were, different learning contracts in terms of the quality of language skills aimed at. Thus fast learners should be encouraged to work for both an accurate and fluent command of the language, thereby utilizing effectively their potential for learning an abstract rule system. For slow learners, on the other hand, it would seem equally advisable to attempt to put into communicative use that amount of language which is within their reach, thereby giving them positive experiences of being able to use the language in a comprehensible way. Communication is thus a relative concept, and everyone should be encouraged to proceed as far as possible within his total learning situation. How far such an individualized approach can be realized in practice depends on the constraints imposed by classroom realities. But it would seem to be a possibility that is worth exploring.

A second important point is the role of affective factors in language use. Language use always involves an element of ambiguity tolerance and risk-taking. Comprehending unheard (and unseen) messages and producing one's own messages requires negotiation skills, perseverance and willingness to attempt communication. Using an acronym, communication could be described as a "SWAP" phenomenon, where S stands for "having something to say", W for "willingness to attempt communication", A for "ability to do so", and P for "personal payoff from so doing". If such attitudinal aspects are to be learned in school, we should be aware of their existence and think of pedagogical ways of developing them in classroom work. This will involve conscious attention to the quality of the learning atmosphere, and it will lead to improvised, personal language use and an extensive treatment of texts, both aural and written. Communicative competence includes an element of communicative self-confidence. This point is also made by Dulay et al., who point out that "All things being equal, the self-confident, secure person is a more successful language learner" (1982:75). A problem with testing is that such attitudinal attributes do not easily lend themselves to objective and reliable measurement. But the problems of quantifiability should not dictate our testing procedures, if we are to evaluate our learners' total communicative skills. As noted above, validity should be given the first priority in testing, while ensuring sufficient reliability at the same time. We need to be critical about the kinds of skills measured by our tests.

An further important point about communicative testing is that it should not undermine the learner's self-confidence and discourage his attempts to use the language. We should therefore try to give him positive feedback about the development of his skills, rather than pinpointing errors in his performance. Concentration on errors will put into focus what the learner does not know, thus leaving aside the important aspects of what he already knows and how much he can communicate with the foreign language. Communicative testing should allow room for both "hard and soft" aspects of evaluation. We must bear in mind that summative tests will inevitably tell learners what we regard as "worth learning" in communicative skills, and they will exercise a powerful backwash effect on teaching. Our tests should encourage communicatively oriented classroom work. In practice we have to make compromises due to existing resources. But such compromises should be made with an awareness of what could be done, and how much of it can be done with available resources.

In mixed-ability groups, tests should be "tailor-made" for the different levels of learner performance. In our approach, the notion of "graded tests" refers to a deliberate and careful design of the items at the hierarchical levels of language processing, ranging from simple recognition to free production. Emphasis has been shifted onto integrative skills, thus enabling learners to show what they can do with their language. A second important point is that learners should be allowed and encouraged to proceed as far as they can. Thus they can attempt to solve any items as far as they wish. This is in accordance with our conviction that we should avoid imposing limits on learner performance. We prefer grading to be tactful, something that takes place automatically in the learner's mind when he sees that he cannot solve the items any further.

Communication is not just a matter of knowing the language system. It is equally importantly a matter of having something to say and the courage to attempt communication, a willingness to engage in discourse. There are numerous possible and acceptable ways of sharing in cross-language communication. It is our task as teachers to help our learners to find their ways and get positive learning experiences from their discovery of the foreign language.

I am grateful to the experimenting teachers at Kaukajärvi and Pikkola Comprehensive Schools (Arja Haavisto, Meeri Halttunen, Ulla Hartiala, Irmeli Kalli, Arja Kauppila, Raisu Kuronen, Ulla Liljeström, Pentti Niemistö, Ulla Nordberg, Ulla Tammi, Riitta Tuukkanen) for lively discussions at several joint seminars. The paper reflects, indeed, the shared insights arrived at in these discussions. I am also grateful to Mr Ian Gurney for comments on this paper, and to Mrs Jaana Rantanen for help with the computer processing of the empirical data.

Bibliography

- Altman, H. and V. James (eds). 1980. Foreign language teaching: meeting individual needs. New York: Pergamon Press.
- Anderson, J.R. 1980. Cognitive psychology and its implications. San Francisco: W.H. Freeman and Company.
- Babad, E.Y. et al. 1982. "Pygmalion, Galatea and the Golem: investigations of biased and unbiased teachers". *Journal of Educational Psychology* 74: 459-74.
- Block, J. (ed.) 1971. *Mastery Learning*. New York: Holt, Rinehart & Winston.
- Bloom, B. 1971a. "Affective consequences of school achievement". In Block, J. (ed.) 1971. 13-28.
- Bloom, B. 1971b. "Mastery learning". In Block, J. (ed.) 1971. 47-63.
- Bloom, B. 1976. *Human characteristics and school learning*. New York: McGraw-Hill.
- Bloom, B. et al. 1981. *Evaluation to improve learning*. New York: McGraw-Hill.
- Carroll, J.B. 1971. "Problems of measurement related to the concept of learning for mastery". In Block, J. (ed.) 1971. 29-46.
- Carver, R.P. 1974. "Two dimensions of tests: psychometric and edumetric". *American Psychologist* 28: 512-18.
- Cziko, G.A. 1981. "Psychometric and edumetric approaches to language testing: implications and applications". *Applied Linguistics* 2: 27-44.
- Dulay, H. et al. 1982. *Language Two*. New York: Oxford University Press.
- Elashof, J.D. and R. Snow. 1971. *Pygmalion reconsidered*. Worthington: C.A. Jones.
- Faure, E. et al. 1972. *Learning to be*. Paris: UNESCO.
- Harding, A. et al. 1980. *Graded objectives in modern languages*. London: Centre for information on language teaching and research.
- Kohonen, V. 1983. "Learners, teachers and graded objectives in communicative language teaching". In Nyyssönen and Mauranen (eds). *AFinLA Yearbook 1983*. 37-67. *AFinLA Series of Publications No 36*. Oulu and Helsinki.
- Kohonen, Viljo. 1984. "On cognitive achievement and self-concept in various learner groups: some empirical findings". In *Differentieringsfrågor i språkundervisningen*. Seminarierapport nr. 11 108-49. Nordiska språk- och informationscentret, Helsingfors.
- Koululait 1983. *Lakikokoelma 13*. Helsinki: Valtion painatuskeskus.

- van der Linden, W. J. 1982. "Criterion-referenced measurement: its main applications, problems, and findings". *Evaluation in Education* 5: 97-118.
- Littlewood, W. 1984. *Foreign and second language learning*. Cambridge: CUP.
- Peruskoulun englannin kielen oppimäärä ja oppimääräsuunnitelma. Kouluhallitus 1982.
- Porter, D. 1983. "Assessing communicative proficiency: the search for validity". In Johnson, K. & D. Porter (eds.), 1983. *Perspectives in communicative language teaching*. New York: Academic Press. 189-209.
- Rask, S. 1982. *Arviointi oppilaan tukena*. Vantaa: Kunnallispaino.
- Ropo, Eero. 1984. *Oppiminen ja oppimisen tyylit*. Acta Universitatis Tamperensis, ser. A vol. 172. Tampere.
- Rosenthal, R. and L. Jacobson. 1968. *Pygmalion in the classroom*. New York: Holt, Rinehart and Winston.
- Rosenthal, R. and D.B. Rubin. 1978. "Interpersonal expectancy effects: the first 345 studies". *The Behavioral and Brain Sciences* 3: 377-415.
- Strevens, P. 1980. "The paradox of individualized instruction: it takes better teachers to focus on the learner". In Altman, H. and V. James (eds) 1980. 17-29.
- Takala, S. 1985. "Criterion-referenced testing in foreign-language teaching". This volume.
- Warries, E. 1982. "Relative measurement and the selective philosophy in education". *Evaluation in Education* 5: 191-203.
- Widdowson, H. 1979. *Explorations in applied linguistics*. Oxford: OUP.

Sample test / English / 7th Grade
Autumn 1983

name _____

Test 1: Fill in the missing words from the box:

Paul and his friend Simon 1) _____ to buy
_____ haluavat (want)
an old boat. Their father can help them to 2) _____
_____ it.
(repair)
They get the 3) _____ paper every day. Today
_____ paikallis (local)
they see an interesting 4) _____ for an old
_____ ilmoitus (advertisement)
rowing boat. They are very 5) _____. It's
_____ kiireisiä (busy)
6) _____ to go to see the boat. It is 7) _____
_____ hauskaa (fun) _____ todella
_____ old, but very 8) _____ so they buy it.
(really) _____ halpa (cheap)

	0	1
1)		
2)		
3)		
4)		
5)		
6)		
7)		
8)		

repair, mean, hurt, loud, reach, bonnet, nearest,
prize, usually, busy, local, game, want, perhaps,
really, interested, fun, problem, harvest, cheap,
expensive, fine, advertisement

Fill in the missing words according to the cue in Finnish:

Sheila Jackson lives in the country. The Jacksons 9) _____
_____ there 10) _____ and 11) _____
_____ viljelevät (grow) vehnää (wheat) kauraa (oats) 11)
There are cows there, too, and Sheila helps to 12) _____
_____ lypsämään
_____ them. Because Roger is the 13) _____
(milk) _____ nopein (quickest)
runner in his school he is often in London in summer. He
usually wins the 100 meters 14) _____. His friend 14)
_____ kilpailu (dash)
Michael always has a good 15) _____ when Roger 15)
_____ istumapaikka (seat)
runs.

	0	1
9)		
10)		
11)		
12)		
13)		
14)		
15)		

Fill in words that suit to the context:

Paul and Simon have parties in an old barn. There is no
16) _____ in the barn, but they have a
17) _____ so they can dance. Paul's
18) _____ music is old rock but Simon is a
heavy rock 19) _____. They 20) _____
their friends to their parties.

	0	1
16)		
17)		
18)		
19)		
20)		

score: _____/20

Test 2: Read the text and underline all verbs in the past tense, both regular and irregular.

Mr Glenn went shopping yesterday. He needed a new coat. 21)
In the shop he saw many nice coats, but they were all quite 22)
expensive. He tried on a dark blue coat, but it looked 23)
too big. The lady showed him a smaller coat and Mr Glenn 24)
liked it a lot. He bought it and drove home. 25)

Look at the pictures and tell what your friend did last summer.



Last summer my friend 1) _____ a good book.
Then she 2) _____ some strawberries and
3) _____ the car. Then 4) _____
5) _____
6) _____

Test 3: Jill tells about herself. Ask her name, age, where she lives, etc. You can also ask her other things than those mentioned in the text. Ask at least 5 questions!

My name is Jill Harper. I'm 12. I live in Hartford. It is a town near London. I've got one brother, his name is Jack. I go to school in Hartford. I like it very much. After school I play tennis with my friend Terry. Terry usually wins, because she is better than me.

1) _____
2) _____
3) _____

0 1

score: ____ / 10

0 1 2

score: ____ / 9

0 1 2

- 4) _____
5) _____

	0	1	2
40)			
41)			

score: ____ / 10

Test 4: John rings Tom. Write down what John says to Tom.

Tom: Brighton 653144

John: Hello, 1) _____ . 42)

täällä puhuu John (this is John speaking) It's about the match.

Tom: oh yes. When is it?

John: 2) _____ . 43)

Tänä iltana seitsemältä (tonight at seven)

Tom: Do you think Bob will win?

John: 3) _____ . 44)

Viime vuonna Philip oli ensimmäinen (last year Philip won)

4) _____ . 45)

Teddy toinen ja Bob viides (and Teddy came second and Bob fifth.)

Tom: But I think he is much better now.

John: Yes, I hope so. 5) _____ . 46)

Tavataan stadionilla (we'll see at the stadion)

score: ____ / 13

Test 5: Tell about last summer's events: where you were and what you did. You can tell, for instance, about travelling, sports or your hobbies or about the weather. Mention at least six events.

- 1) _____ 47)
2) _____ 48)
3) _____ 49)
4) _____ 50)
5) _____ 51)
6) _____ 52)

	0	1	2	3
47)				
48)				
49)				
50)				
51)				
52)				

score: ____ / 13

Test 6: Read the text carefully and then answer the questions in Finnish.

Sport is a very important part of life in an English school. Nearly all schools have their own sports fields. Each school has its own teams, and they play against the teams of other schools. Most games between schools are on Saturdays. Many of the pupils who are not playing come to watch their school team.

In winter boys play football or rugby. Rugby is a very hard game, in which the players carry the ball in their hands. There are many famous football clubs in Britain, and even people in other countries follow their favourite English team.

If you go to England in summer, you can see men in white clothes playing on a big round field. They use a thick wooden bat to hit a small red ball. This game is cricket. Nearly every village has its own cricket team, and there are matches every weekend. To foreigners cricket seems a slow game - an international match can take five days.

0 1 2

- 1) Mitä kaikilla englantilaisilla kouluilla on?
(what have all English schools got?)

53)

- 2) Milloin ottelut yleensä pelataan?
(when are games usually played?)

54)

- 3) Mitä ne oppilaat tekevät, jotka eivät pelaa?
(What do those pupils do who do not play?)

55)

- 4) Miten palloa kuljetetaan rugbyssä?
(How is the ball carried in rugby?)

56)

- 5) Millainen peliasu on kriketinpelaajilla?
(what do cricket players wear?)

57)

Your total score: _____ / 90

score: _____ / 10

Appendix 2.

Item Analysis

ITEM No. Item Facility
corr. Value (%)

TEST N.O 1 (Vocabulary)

1/	1	0.67	80.77
2/	2	0.73	81.73
3/	3	0.68	85.58
4/	4	0.55	84.62
5/	5	0.68	75.00
6/	6	0.53	72.12
7/	7	0.64	77.88
8/	8	0.68	74.04
9/	9	0.73	68.23
10/	10	0.84	60.58
11/	11	0.78	68.23
12/	12	0.62	77.88
13/	13	0.63	74.04
14/	14	0.68	74.04
15/	15	0.72	73.08
16/	16	0.74	58.62
17/	17	0.72	60.58
18/	18	0.46	36.54
19/	19	0.51	45.18
20/	20	0.45	31.73

Mean 0.65 68.17
s 0.10 14.73

TEST N.O 2 (Past tense)

1/	21	0.60	86.54
2/	22	0.58	91.35
3/	23	0.68	74.04
4/	24	0.53	58.62
5/	25	0.52	81.73
6/	26	0.56	88.46
7/	27	0.56	88.46
8/	28	0.58	84.62
9/	29	0.62	82.69
10/	30	0.57	75.86
11/	31	0.42	61.54
12/	32	0.63	77.88
13/	33	0.62	80.77
14/	34	0.78	66.35
15/	35	0.71	56.25
16/	36	0.70	66.83

Mean 0.60 76.44
s 0.08 10.84

TEST N.O 3 (Cued Interview)

1/	37	0.73	75.96
2/	38	0.66	78.85
3/	39	0.71	71.15
4/	40	0.75	65.87
5/	41	0.67	63.84

Mean 0.70 71.15
s 0.03 5.70

TEST N.O 4 (Dialogue compl.)

1/	42	0.66	81.73
2/	43	0.66	63.78
3/	44	0.76	75.00
4/	45	0.57	83.65
5/	46	0.52	28.85

Mean 0.63 66.60
s 0.08 20.12

TEST N.O 5 (Past Events)

1/	47	0.77	61.22
2/	48	0.80	57.37
3/	49	0.78	60.58
4/	50	0.78	54.17
5/	51	0.86	51.82
6/	52	0.77	49.36

Mean 0.80 55.77
s 0.03 4.36

TEST N.O 6 (Reading Compr.)

1/	53	0.32	44.23
2/	54	0.44	84.62
3/	55	0.50	79.81
4/	56	0.26	81.35
5/	57	0.33	63.46

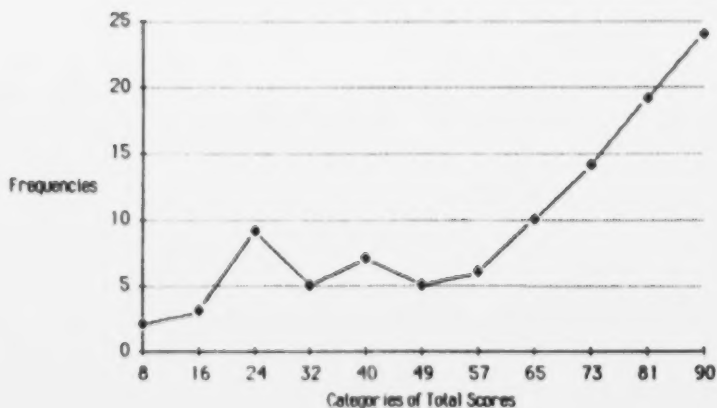
Mean 0.41 72.68
s 0.10 18.85

Correlations Between Tests (7 = Total Scores)

	1	2	3	4	5	6
2	0.904					
3	0.844	0.851				
4	0.646	0.817	0.792			
5	0.842	0.866	0.773	0.799		
6	0.580	0.567	0.523	0.503	0.514	
7	0.958	0.955	0.894	0.895	0.925	0.651

Test	Items	Mean	s	Reliability
1	20	13.635	6.097	0.942
2	16	14.125	5.156	0.902
3	5	7.115	3.017	0.875
4	5	8.337	3.362	0.822
5	6	10.038	6.005	0.930
6	5	7.269	2.661	0.651
7		60.519	23.801	0.977

Frequency Distribution of Total Scores



ON THE WHYS AND HOWS OF LANGUAGE TESTING

Torsten Lindblad (University of Gothenburg)

These reflections on why we test and how we test our students' command of a foreign language are not meant to be the all-embracing treatment on the somewhat grand scale that the title of the paper might suggest. Hopefully it will be, however, a reminder of the fact that our tests must never become ends in themselves but must always remain means to different ends. This is, at bottom, one way of approaching the problem of the validity of foreign-language tests.

1. Background

When people disagree in their opinions as to what is or what is not a good language test, the "deep structure" of the disagreement might be found in differences concerning certain underlying, basic concepts rather than in the "surface structure" of test format and item content. Let us, therefore, briefly mention six background concepts which might be useful in starting a discussion of what is often, implicitly, taken for granted. For reasons of space I shall have to limit myself to hinting at what is meant rather than treating in full each point mentioned.

1. View of knowledge. What is it a student "knows" when he has acquired command of a foreign language? What can be measured in quantitative terms and what must be estimated in qualitative ways? Is there a risk that too much attention is paid to what can be calculated in figures, as means and standard deviations etc? Also: If we say - as is often done - that a student knows a certain number of words in the foreign language, what exactly do we mean by "know" and what is "one word"? Is 'have' one word or three words in 'I have a dog', 'He will have it that' and 'You've been had'? And furthermore: In formulating the goals for the weakest students, i.e. in deciding on the basics of a foreign language, do we use a list of one hundred frequent words or do we mention a dozen important situations in which the student is supposed to be able to cope?

2. View of language teaching and learning. What does it mean to learn a foreign language? What is the role of the four skills as compared to attitudes to the foreign country and to speakers of the language and, for that matter, to speakers of foreign languages in general? And what attitudes to themselves and their own ability do our students acquire? Will they dare and be willing to use the foreign language when the situation arises? Is it, in other words, only new words and grammar paradigms that count?

3. View of language. What language do we teach, what language norm do we have, implicitly or explicitly? Where do we "draw the line" in accepting student responses in open-ended tests? Which of the following three sentences could be given points and what principles could be

formulated in order not to leave us in an *ad hoc* swamp: 'There's lots more questions I'd like to ask', 'There ain't gonna be any room for us', 'He said he don't like it'. I would suggest that comprehensibility is not enough (third example not accepted), usage will have to depend on the context (second example accepted in dialogue but not in written discourse) and suitability, a rather strict demand, well in accord with the concept of 'communicative competence', will have to be judged with the help of native informants (the first example fully acceptable today, whatever our grammar books might still say). For our purposes we might do well to use a prescriptive view in language teaching but a rather more descriptive one in testing. (These ideas are worked out in more detail in a paper, "The Language Norm - Does it Exist in English", by Professor Jan Svartvik, which can be ordered from the author of this paper.)

4. View of language skills. What is the relationship between linguistic competence and communicative competence? What are the relationships between the various factors of the foreign-language proficiency of the students? This is, of course, closely related to the view of knowledge and the problems mentioned under that heading.

5. View of language testing. Do we have (with Stevenson's terms) a restricted or an expanded view of language tests and testing? What are the risks in using one kind of test or another? Do the three well-known trends formulated by Spolsky - originally referred to as prescientific, structuralist-psychometric, and communicative - reflect swings of the pendulum between what C.P. Snow called "the two cultures"? If so, do we favour a communication-oriented, humanistic approach that "dares" to take into account aspects of language proficiency that cannot be objectively scored? How do we strike the happy medium between quantitative and qualitative approaches? Moreover: What do we want to test - what the students know, what they do not know, or what they should know?

6. View of students (human beings in general). Is the student the object of our teaching efforts or the subject responsible for learning the language? Do all students have the same ability to learn a foreign language? If not, does that influence our view of their "value" in any respect? Do all students have the same right to have the feeling that they have succeeded in some way? How do we handle the affective effects of our tests?

The way we answer all of the above questions will influence our views on the "classical" testing problems of validity, reliability, backwash effects, marking principles, and use of test results.

2. Why do we give tests?

Tests are often given, quite naturally, as part of the language course without too much consideration of questions like: In whose interest do we give this test? and What do we want to know something about? I would suggest that tests can be given for the following ten purposes.

- | | |
|---------------------------------|---|
| 1. Diagnosis | what do the students not know? |
| 2. Resources | who needs most help? |
| 3. Selection | who knows most? |
| 4. Prognosis/Guidance | who is most likely to succeed? |
| 5. Examination | do the students know what they should know? |
| 6. Grading | who knows most/the least etc? |
| 7. Evaluation | what have we taught the students/
have they learnt? |
| 8. Influence
(Backwash) | what the students should know and how,
what the teachers should test and how |
| 9. Stimulation
(Information) | what they know/do not know - but should
know |
| 10. Research | what is the object of study, comparisons
etc |

Tests produced for one purpose can not, as a rule, be used for a different one. This is often overlooked. When the validity of a test is commented on, one must know why it was constructed the way it was. What might be quite a good test for one purpose can be invalid for another. The term "diagnostic test" is quite often, in my experience, misapplied to almost any test. Diagnostic tests are primarily meant to find out what the students do not know, what needs to be taught or practised more.

Tests to determine the distribution of resources are used in Sweden in the teaching of Swedish as a second language to immigrant students. There are limited resources available and one needs to know who can join a regular class and who needs special teaching etc. This will be referred to in more detail below.

Some of the reasons for giving tests are more relevant in certain countries. In Sweden, for example, we give standardized tests on a nationwide basis for grading purposes, but they are not public examinations in the British or Dutch sense. Tests for evaluation come close to those used for research purposes. The American National Assessment and the British Assessment of Performance Unit projects are examples of evaluation studies: one wants to "take the temperature" of the conditions in the country.

Almost any test has some kind of backwash effect. The teacher who gives tests at regular intervals will tell her students, whether she realizes it or not, what she considers important, in what way the students are expected to use their knowledge etc. Nationwide tests given by some kind of examination board or other central body will similarly influence both students and teachers. This being so, care must be taken that these effects are positive rather than negative. This means that the backwash effects become not just chance influences that can not be avoided but purposes of their own. One reason (but not the only one, of course) why we consider the listening comprehension test so important in the Swedish standardized tests is that we believe teachers pay more attention to listening exercises because of this test than they would otherwise do.

In the following section I shall describe three different Swedish tests or test batteries, given on the basis of different "whys". In describing them I will make some further comments in detail on various background reasons for giving tests.

3. The Hows of Language Testing - three examples

Aspects to be considered under the heading "how" could refer to such things as test format, test content, marking principles, and use of test results. What weight do we attribute to validity as opposed to reliability, the two sometimes tending to counteract each other? What weight do we give various parts of the overall command of the language? How important is the actual content of the texts used? There may be a risk that language is considered a sufficient end in itself so that the texts become boring even though they are linguistically acceptable. To what extent do we use multiple-choice formats as opposed to open-ended items or student production on a larger scale? What principles do we use in marking essays etc? And so on, and so forth.

3.1 The "SVAN" Tests. More than 10 % of the students in Swedish schools are of immigrant background, many of them being bilingual and having Swedish as their second language. It has been found over the years that many - but, of course, by no means all - of these students meet with difficulties in many subjects because of language deficiencies. These are often detected too late, and there has been a general feeling that something ought to be done to "find" these students early in their school careers so that extra teaching efforts could be made. Within a research project at Gothenburg university, headed by the writer of this paper, a series of tests were developed a couple of years ago, the so-called SVAN tests (svenska som andraspråk, i.e. Swedish as a Second Language).

Why did we need these tests? There are 100,000 bilingual students in the comprehensive school. Instruction in Swedish as a Foreign Language (SFL) is compulsory for those students judged to be in need of such instruction. Resources being limited, the teachers have to find the students who need help the most. Any experienced teacher will be able to tell the beginner from an intermediate student, but we have found over the years that it is extremely difficult even for a well-trained teacher to tell the student with good fluency and pronunciation but a limited vocabulary from the one who has a truly native-like command of the full language. The superficial ease of the former tends to hide the limitations in the "depth" compared with the latter.

So we had to try to produce difficult tests - at each age level - to detect differences between bilingual students and monolingual students of the same age and cultural background. We also faced - but I am not sure that we solved - the problem of differentiating between general intelligence and linguistic proficiency: even among monolingual speakers there are great differences, of course, due to variation in general ability. We also wanted, through our tests, to influence the teaching procedures favorably, i.e. we aimed at positive backwash effects. We know that many, perhaps most, of the SFL teachers are not specialist language teachers. The tests were seen as part of an in-service training program.

There are oral tests for each of grades 1, 2, and 3, and both oral and written tests for grades 4, 6, 7, and 9 of the comprehensive school, students aged from 7 to 16. (It might be added that there are similar tests in Finnish, Serbo-Croatian, Croatian-Serbian, Spanish, and Turkish for the testing of the "home languages" of the bilingual children, as well as a set of models for the testing of other "home languages".)

The tests are based on pictures and series of pictures and on meaningful pieces of texts. The testing of an active and a passive vocabulary is an important part of the test battery. The words have been chosen from different spheres of life, the most important perhaps being the school environment, including textbooks, that we know from experience cause the bilingual child great problems.

The teachers were advised to evaluate pronunciation, vocabulary, grammar, and fluency, as compared to "a comparable (age, region, social background etc) monolingual student with Swedish as the first language". This was done at six levels from F = No difference to A = No proficiency.

The tests have been used on a national level for two years now and seem to work well. One of the main effects of using these tests has been that many students who were previously considered to have a proficiency at level F, i.e. no difference as compared to a monolingual student, have been given an E or a D. This in turn means that these students do get that extra SFL teaching that they still need rather than being put in the regular Swedish class. We shall have to wait for a number of years before we can tell whether there will be fewer problems at higher levels in the school system and whether the bilingual students do better in the more difficult lines of study.

The tests might be considered to be too "structuralist-psychometric" in that so much stress is put on vocabulary. The words are always tested in natural "communicative" contexts, however, and represent a great step forward from the all-too-common overemphasis on correct grammar. One of the main "whys" of the tests being to detect the very good from the truly native-like, it has been necessary to use strict and reliable techniques. This is an example of the fact that the "whys" explain the "hows" - but that, of course, does not mean that what has been done is not open to criticism and differences of opinion as to the results.

3.2 The Standardized Tests. In the Swedish school system nationwide standardized tests are given in certain subjects, among them foreign languages: English in grade 8 of the comprehensive school and English, French, and German in grade 2 of the *gymnasieskola*, corresponding to the eleventh year of school. These tests are given for one main reason: to help teachers give their students fair grades. The five-step grading system being norm-referenced (7-24-38-24-7 per cent given each grade), the students of the individual class have to be compared to all students in the country somehow. Since the tests are nationwide and set by the National Board of Education, they tend to become a concretization of the national curriculum, and the backwash effects are great and have to be taken into consideration in constructing the tests. This refers both to what is being tested and how it is done.

In the tests given in the spring term every year at *gymnasium* level there are the following parts: listening comprehension (multiple choice), reading comprehension (multiple choice), vocabulary-grammar tests (both MCT and open-ended), and "integrative tests" (i.e. cloze). In French there is also a dictation test. In the last few years there has also been "free written production" (see below). The listening tests have sometimes consisted of actual, live recordings, but usually we have worked with scripted dialogues recorded by professional actors,

sounding somewhat like radio plays. This year's English test was a heated discussion in a miner's home on an on-going strike, originally prepared long before the present strike in Britain started. It was thus very topical and quite life-like. The reading tests have consisted both of longer pieces of texts with many questions and shorter ones with just one question for each extract. The "vocabulary-grammar" items are always monolingual and contextualized. Quite often the student is asked to produce a full phrase, and the technique comes close to a "notional-functional" approach or, if you will, a communicative test. In discussing different tests it is well to bear in mind that different "names" can well stand for quite similar tests, whereas the same "name" might hide great differences.

There is no oral test in the standardized test battery and the written essay (see below) is still included on an experimental basis. It has been found, however, in our own studies as well as in many others that the correlation between, for example, teachers' grades (which do include oral proficiency) and the standardized tests is quite high. What the teacher needs to know is the general level of her class and the approximate number of grades at each level. We can tell from our studies, for example, that teachers tend to give girls too high grades even though the girls do score somewhat higher on the standardized tests also. They also give students in slower streams lower grades than in more advanced streams even though they should be graded separately and not compared. Students in academic, "theoretical" lines of the *gymnasieskola* also tend to be treated more leniently than many of those in "practical" lines. Injustice is also sometimes found when students from different social backgrounds are compared. All these variations in grading which are not related to actual proficiency are counteracted by the national standardized tests.

The strong stress on communication-oriented techniques in the reading and listening tests as well as in the "vocabulary-grammar" parts tends to have backwash effects on the teaching and testing of languages in our schools. Translation is relatively little used and strict grammar tests with little or no communicative content are probably less frequent than they would be if there were no standardized tests. Since it is compulsory for teachers to use most of the tests, we try not to go against teacher opinions, and to a certain extent the tests are compromises, trying to "push" in certain directions, however.

These tests would not be very useful for diagnostic purposes, nor are they meant to be used for an evaluation of educational standards (even though certain conclusions can be drawn from a close scrutiny of the results). In evaluating the "hows" one should, again, bear in mind the "whys": primarily to grade students in a norm-referenced system and, secondly, to show how the National Board responsible for the nationwide curriculum wants it to be understood.

3.3 "Free written production". It is becoming increasingly common in many countries to put more stress on oral and written production even though the techniques used must be more or less subjectively scored. Why do we need such tests? Don't the ordinary written tests give a good enough picture of the levels of proficiency of our students? Even if foreign-language proficiency is not one unitary factor but

consists of a number of more or less distinct traits, are not the intercorrelations high enough and the test batteries we use varied enough to account for almost all of the variance?

Even if the answer to the above questions is, in all likelihood, "Yes", we still have the backwash effects, the influence of our tests on students and on teachers. It is no good stressing in the curriculum that the goal is, among other things, command of the spoken language and an ability to express oneself in writing - and then not testing it. Well, teachers are supposed to do this on their own. But we need to give them models, and the teachers have to know whether their own students do better at these things than other students. It is quite common for teachers to say that, true enough, their students scored a certain mean on the standardized test, but then their students are so good at speaking the language that the teacher has to use a higher mean in her final grades.

In order to meet this situation we have started, in the last few years, to produce tests to evaluate free (well, more or less so) written production. (Models for the testing of oral production have been produced in French and German, and similar tests have been used in English as well, but these have not been followed up quite as carefully and will not be discussed here.) There are two problems that have to be faced in this context: how do we make students produce, and how do we mark what they produce?

Experiments have shown that one and the same student can produce next to nothing if he is given a topic to write on that does not interest him at all, or he can produce a long and interesting tale on something that gets his imagination going. If we are primarily interested in his command of the foreign language, we must see to it that he is motivated to use so much of it that we have a sample to evaluate. We must also consider the "naturalness" of the task. For what purpose is a Swedish student ever likely to write something in English, French, or German? The obvious answer is: letters. We have tried other tasks, but various kinds of letters have proved suitable for our purposes. We had a native speaker write a fictitious letter-to-the-editor criticizing Swedish young people for bad and irritating behavior. The students were then asked to write their own letter to contradict or agree with "Mr. Brown". Many of these letters were rather heated and the writers may very well have forgotten that they were taking a test.

In French the first experiments with written production are being made this year. It was feared that French would be too difficult so that nothing would come of it. The kind of *précis* model used seems to have worked well, however. The students were given a text to read. It was then collected and they were asked to retell the story in a different tense. The results are quite promising. It should be borne in mind that while English is taken by 100 % of all students, less than 10 % take French, usually the linguistically very best ones.

To arrive at some kind of nationwide agreement as to the marking of the essays, we have produced booklets of examples. The papers are attributed to three "lots" of roughly equal size: average, below

average, above average. A group of teachers and teacher trainers read a couple of hundred randomly-selected papers and grade them. A dozen of them are included in the booklet and commented on. Reasons are given for the grades given. These include linguistic merits and mistakes as well as the treatment of the subject and the arrangement. Most examples illustrate the "borderlines" but there are also very good and very poor essays as well as "extremely average" ones. This technique has been well received by most teachers.

The grading is thus a combination of norm- and criterion-referencing, and communicative content is usually stressed more than formal correctness. Thus both the kind of task we set and also the treatment of the products constitute the backwash effects.

The standardized test in English in 1984 correlated .75 with teacher-given grades (given two months before the test) and .69 with the free written production test. The latter correlated .71 with teachers' grades. These figures indicate that there is great agreement between the different measures of student proficiency even though examples can also be found of students who score very high in the written essay test and low on the standardized test and vice versa.

4. Conclusion

I have tried to discuss the importance of analysing why we want to give a test before we start constructing it as well as of knowing why a certain test was constructed the way it was before one starts criticizing it. Tests can be made in many different ways but these "hows" should always be related to well-defined "whys". We should also remember that however important language testing is, language teaching is always more important. The most important of all, however, is language learning, for without it no testing is meaningful at all. Language tests, thus, are only means to certain ends and must never become ends in themselves. We must, on the other hand, always keep the ends in mind, which might be all the more important for us who are sometimes referred to, and perhaps refer to ourselves, as the language testing circles.

CURRENT TRENDS IN FL TESTING IN CZECHOSLOVAKIA AND THE SOVIET UNION

Jan Prucha (Comenius Institute of Pedagogy, Praha)

In the present paper I will deal with two issues. In the first, more general part I will explain how FL testing has been interpreted in present-day pedagogical and applied linguistic theory in Czechoslovakia and the Soviet Union. In the second part I will describe the main tendencies in the application of FL tests with respect to some important empirical investigations.

1 The general situation

The general situation in FL testing in Czechoslovakia and the Soviet Union is rather different from the situation in most West European countries. Though FL testing is nowadays recognized as an important means of assessing FL achievements, in school practice FL tests are used quite rarely. In schools of different levels it is mainly traditional methods of evaluation (i.e. oral and written examinations) of FL achievements that are applied.

However, new developments have been taking place in socialist pedagogical theory in the last few years:

(1) There is a growing trend towards increasing the efficiency of the instructional process in school and of its results. Educational theorists in Czechoslovakia and the USSR now emphasize the necessity for pedagogical research to become more efficient.1) Consequently, they also postulate an important role for pedagogical diagnostics. This requires, among other things, the development of more precise and objective methods which could be applied in the evaluation of the real process of instruction and real outcomes of that process. For example, some researchers strongly argue for the reinforcement of the role of exact measurement of pedagogical objectives and processes.2) Thus, socialist pedagogical theory does not reject tests and testing as an instrument of scientific investigation. On the other hand, however, it has a critical attitude towards their abuse or the overestimating of their application: "Tests serve as only one of the measuring procedures in the framework of the whole system of research methods" (Skalkova 1983: 103). Didactic tests (including FL tests) are therefore considered to be, first of all, a means of educational research and not primarily a procedure applied by teachers in the general everyday practice of school instruction.

(2) In consequence, the most widely used means of evaluation of educational outcomes in school practice are not tests (in the sense of standardized, objective tests) but traditional methods of examination, i.e. oral or written tests. This kind of examination uses marks (a 5-point scale in both Czechoslovakia and the USSR). However, such evaluation of pupils' achievements has been subjected to strong criticism by some pedagogical and psychological researchers in Czechoslovakia and the Soviet Union. The critics present serious empirical evidence concerning mainly the high subjectivity of traditional marking: e.g., one and the same performance of a pupil is marked very differently by a group of teachers or even by the same teacher after some time has passed. Besides the subjectivity, there are other shortcomings for which traditional examinations have also frequently been criticized, such as low economy, poor comparability, a strong dependency of marking on teacher's personal preferences, etc.)

On the other hand, however, socialist pedagogy has also stressed some important positive features of traditional examinations which are lacking in tests. The following table gives a comparison of examinations and tests.

Table 1

Tests vs. individual evaluation (marking):
Comparison of positive and negative features

	individual evaluation (FL)	test (FL)
+	- possibility to approximate to real communicative situations - possibility for teacher to put additional questions, tasks - possibility for teacher to recognize pupil's motivation, endeavour, etc. - not requiring special preparation by teacher - possibility for pupil to be creative - possibility for pupil to be helped (guided) by teacher - during oral evaluation pupil learns to speak in public setting - evaluation is integrated - contributes to developing good social relations between teachers and pupils	- high objectivity of testing - less time-consuming - possibility to compare results - high precision, reliability, good discrimination, etc.

- subjectivity of marking - time-consuming - no fixed criteria for comparison	- artificial activity for pupil - requiring mostly mechanical recall of knowledge - not reflecting pupil's attitudes, opinions, etc.
---	--

(3) During the last few years efforts have been made in socialist pedagogy to eliminate the most serious shortcomings of the traditional exam, viz. low objectivity. In the field of FL teaching in the USSR this tendency has been realized mainly through criteria for evaluating pupils' knowledge and skills.4)

An example of the criteria is given in Table 2.

Table 2

Criteria for evaluating pupils' FL achievement
(Soviet system)

Generally, marks are awarded 4 times per school year, with an overall mark at the end of the year in figures: 5 - 1 (5 being the highest mark)

Marks in FL:

a) separate mark for each of 3 FL skills:

- oral production
- aural comprehension
- reading comprehension

example of criteria for evaluating pupil's oral speech:

mark 4: if a pupil expresses his thoughts in the FL with only non-significant deviations from the correct usage (e.g. articles in English, or prepositions) but in other respects his performance is in correspondence with goals defined by the syllabus

mark 3: the same as for mark 4 with more significant deviations from the correct usage (not, however, preventing understanding)

mark 2: with such deviations which prevent understanding

b) overall mark for FL competency: developed as an integration of 3 practical evaluation

example: an overall mark of 4 is attained by a pupil if

- his performance in all 3 partial competencies has been evaluated as 4,

- or his oral competence is 4, but two others are 5,
- or his oral competence is 5, two others are 4 or at least 3

(Source: Normy ocenki znanij, umenij i navykov ucascichsja po inostrannym jazykam (Criteria for evaluation of knowledge, skills and habits of pupils in foreign languages). In: Inostrannyye jazyki v skole, 1978, No. 1, pp. 3-6.)

Of course, the criteria for evaluation have not been intended to replace the use of FL tests (cf. section 2).

(4) In Czechoslovakia one can find another way of increasing the objectivity of traditional evaluation of pupils' FL achievements. Starting in 1984, new curricula have been implemented in our gymnasiums (i.e. upper secondary schools). As regards FL the new syllabuses have been constructed according to the principles of the "threshold level" for learning languages in schools.⁵) This means that the syllabuses define desirable learning outcomes of FLs (in Czechoslovakia: Russian, English, German, French) in terms of concrete objectives and the respective communicative functions and situations. This specification makes it possible to assess more accurately the extent to which the objectives have been achieved by learners.

For example, the new syllabus for English in grades 1 - 2 of the gymnasium⁶) specifies 27 basic communicative situations and 35 language functions which the learners should acquire. However, so far no inventory of grammatical and lexical means for the given purposes has been specified (not, that is, in such detail as has been done in van Ek's concretization). It is only stated that gymnasium learners have to achieve active knowledge of some 500 English words per grade, i.e. about 2,000 English words as a whole.

Unfortunately, so far there have not been standardized FL tests corresponding to the new syllabuses. But I have to state here that the conditions for applying such tests are very favourable in Czechoslovakia. We have a centralized system of creating and approving all syllabuses and textbooks. This means that all learners in our gymnasiums learn English or Russian on the basis of the same curricula and textbooks. This makes it possible to apply FL tests to large-scale populations without any disturbing impact of different textbooks as is the case in non-socialist countries. But as pointed out above, until recently FL tests have been used mainly for purposes of research and only rarely as practical instruments of evaluation in schools.

2 FL tests applied for research purposes

In the Soviet Union the theory and application of various FL tests is dealt with in several research centres. The most important among them are

- a) Moskovskij gosudarstvennyj pedagogičeskij institut inostrannykh jazykov im. M. Torejza (Moscow State Pedagogical Institute of Foreign Languages), i.e. a special university preparing Soviet translators and interpreters.
- b) Institute russkogo jazyka im. A.S. Puškina (Puškin's Institute of Russian Language, in Moscow), i.e. the institute serving the purposes of teaching Russian to foreign students and specialists who study in the USSR.
- c) The most intensive work in the theory of FL testing has been carried out in Estonia, in the pedagogical and psychological centres of the universities in Tallin and Tartu.7)

Soviet findings in FL testing are published mainly in the journals *Inostrannyje jazyki v škole* (Foreign Languages in School) and *Voprosy psichologii* (Problems of Psychology).

In Czechoslovakia there is no special centre for research on FL testing. However, since 1968 a state establishment called *Psychodiagnostické a didaktické testy* (Psychodiagnostic and Didactic Test), in Bratislava, has published various tests (among them standardized proficiency tests for Russian at the basic school level). This is an establishment with the monopoly of creating and publishing tests for Czechoslovakia.

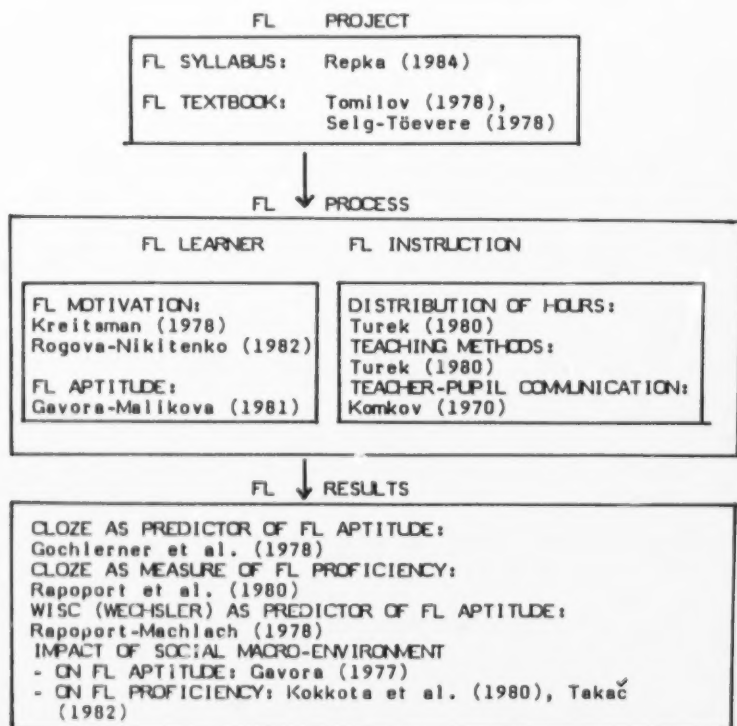
In what follows I try to present a brief survey of investigations in which FL tests have been applied as an instrument of explanation of some components of FL learning-teaching. The process called FL learning-teaching is explained in socialist pedagogy and linguo-didactics as a system. The system consists of its components and is ruled by its determinants. Therefore, my question was simply not how FL tests have been applied in research as such but how FL tests have been used as explanatory tools for some components and determinants of the system called FL learning-teaching.

To give a concrete example: FL tests - as they have usually been applied - measure certain results (achievements, proficiency) of FL teaching. They do not normally assess the causes which account for these results. If a group of learners has, according to the test results, a

given FL proficiency, what are the real causes which produce the assessed state? How have the results been brought about? Is it, for example, due to the distribution of teaching hours for FL in particular grades of school? How do the textbooks contribute to the outcomes, etc.?

On the whole, one has to ask what causes and impacts determine the concrete results of FL teaching at a certain level. For this purpose I have examined how FL tests have been used in Czechoslovakia and the Soviet Union. What I have found is presented in the following figure showing how various levels of FL reaching have been covered by testing.

Fig. 1: Levels of FL teaching: Coverage by tests



I shall now present a very brief commentary to some of the most interesting investigations.

(a) The work of the Soviet researcher Tomilov (1978) is an example of an investigation indicating that the results of FL tests are caused by certain qualities of textbooks.

Table 3

Characteristics of the vocabulary in English textbooks in the USSR (grades 5 - 8)

1. Quantity of vocabulary

	Soviet republics			
	Armenia	Estonia	Georgia	Russia
V	1,029	1,153	1,293	974
T	297	420	333	420
Q	3.5	2.7	3.9	2.3

V = number of lexical units (types) of the "active vocabulary" in textbooks

T = number of teaching hours (grades 5 - 8) for English

Q = average lexical loading per 1 hour ($Q = \frac{V}{T}$)

2. Quality of vocabulary

V (total amount of active vocabulary in English textbooks in 8 Soviet republics): 2,580 lexical units

V (lexical units common in all textbooks): 270 lexical units (= 10.5 % of V)

3. Frequency of vocabulary

Only 62 % of V (= 2,580) belong to the 3,000 most frequent words of English (as counted by Thorndike-Lorge).

The author comes to the conclusion (p. 51) that the school leavers of basic schools in 8 Soviet republics have acquired different English vocabulary and can thus mutually communicate in English by using only 270 words common to all of them. Consequently, English proficiency tests cannot be applied for interrepublic comparisons in the Soviet Union.

(b) Another example concerns the matter of what the effect is of the distribution of teaching hours on FL achievements in pupils. In Czechoslovakia, Turek (1980-81) has applied a special proficiency test (in German) to two groups of pupils in secondary economic schools (grade 1):

group (A) which had 3 hours of German per week,

group (B) which had 2 hours of German per week.

Both the groups were taught according to the same curricula and the same textbooks; the only variable was the amount of time devoted to teaching. The hypothesis was that the proficiency of group (A) would be significantly better, especially in skills. A special test was applied assessing pupils' grammatical knowledge, reading comprehension and written production in German.

The main finding was that the hypothesis was not proved: the proficiency of group (A) was not significantly better than of group (B). One can thus conclude that (on the condition that the test concerned is valid and has discriminative power) a test can be used as an instrument indicating how FL proficiency of learners is (or is not) influenced by the number of teaching hours.

(c) A further example concerns the investigation of FL aptitude: In Czechoslovakia, Gavora and Malikova (1981-82) and Gavora (1977) have contributed to solving the question of what are the determinants of FL aptitude in pupils in basic schools. The research of Gavora and Malikova (1981-82) was carried out on a population of 455 pupils in the 4th, 6th and 8th grades of basic schools in Slovakia. Three determinants of FL aptitude were hypothetically expected as the most relevant: age of pupils; sex of pupils; their socio-economic background. A special FL aptitude test was applied consisting of 4 subtests:

- a grammatical subtest, assessing ability to perform perception and discrimination of words in an artificial language,
- a phonetic subtest, assessing ability for aural perception and discrimination of words in an artificial language,
- a lexical subtest, assessing the learners' LT memory for words in their mother language (Slovak),
- a special subtest assessing the ability to learn words of an unknown foreign language. It is perhaps interesting here to mention that Finnish words were used as representing words of an "unknown foreign language" and the task of the pupils was to learn 20 Finnish words (presented with their Slovakian equivalents) and after 8 minutes of learning to recall them (e.g. lapsi = child, oppia = ucit'sa, yksi = jeden, etc.).

Briefly, the findings were as follows:

The ontogenesis of FL aptitude depends significantly on the age level of learners: the higher the age of learners the more developed is their FL aptitude. Thus, the widespread opinion that younger children have better FL aptitude than older children was not confirmed (in conditions, of course, of learning in school, not in natural conditions as, for example, in a bilingual family). As for inter-sexual difference, it was found again that girls have significantly better FL aptitude than boys (in all 3 age groups) though there was no significant difference in IQ level between the sexes. A significant determinant of FL aptitude is the socio-economic background of learners - both the social micro-environment (profession and educational level of parents) and the social macro-level (type of locality the pupils live in).8)

(d) The last example I would like to mention here is based on investigations of Estonian colleagues. As already said, it is in Estonia that perhaps the most intensive work on FL testing has been done as far as the Soviet Union is concerned. There are two reasons for this: Estonian people have a great need to learn foreign languages, particularly Russian (as a language of communication between Soviet nations), German (as traditionally learned by Estonians) and English (in the last few decades). Another reason is that Estonian specialists in psycholinguistics and educational psychology are well acquainted with latest developments in FL testing.

I have found the following main tendencies in Estonian research on FL testing:

Much effort has been devoted to exploring the relations between FL tests and tests measuring "intellectual abilities" in pupils.9) Here mainly WISC tests (Wechsler Intelligence Scale for Children), especially its verbal sub-tests, have been compared with results of some didactic (proficiency) tests for English and German. For example, Rapoport and Machlach (1978) have found a highly significant correlation between verbal WISC and FL proficiency tests in several independent experiments and, at the same time, a relatively low correlation between WISC and marks awarded to pupils through traditional FL examination as well as evaluations given by expert groups. Consequently, the authors argue that verbal WISC is a reliable predictor of the ability of Estonian pupils to acquire foreign languages taught in schools.

The same endeavour has been devoted to investigating cloze procedure as a tool of language testing. It is interesting that Estonian specialists have come to the same conclusions as Enkvist and Kohonen in Finland:10) Namely, that cloze tests have considerable advantages in comparison not only with traditional exams (marking) but also with multiple choice tests; and, further, that cloze can measure integrative language skills.

For example, Gochlerner, Rapoport and Sotter (1978) found that cloze (applied on texts in the mother language, i.e. Estonian and Ukrainian) reveals a high and significant correlation with proficiency in English (measured by a battery of proficiency tests). The testing was done in groups of learners of grade 10 (i.e. age 17) of Soviet secondary school (cf. Table 4).

Table 4
Correlation of FL proficiency with cloze (M.)
and verbal abilities (M.)

cloze (Estonian, Ukrainian texts)	/	FL (English) proficiency measured by battery of didactic tests	corr.:	sign.:
			0.795	p (0.005)
verbal ST memory for / words (listening)		"	0.306	p (0.05)
verbal ST memory for / sentences (listening)		"	0.399	p (0.025)
verbal ST memory for / words (reading)		"	0.362	p (0.025)
verbal ST memory for / sentences (reading)		"	0.335	p (0.05)
verbal ST memory for / texts (reading)		"	0.478	p (0.005)
verbal LT memory for / texts (reading)		"	0.631	p (0.005)

(Source: Gochlerner, Rapoport, Sotter (1978).)

Another type of investigation has been concerned with how cloze can measure productive skills in a foreign language. I shall summarize here an extensive investigation made by Rapoport et al (1980). The investigation aimed at the verification of the following hypothesis: cloze provides a characteristic picture of the real state not only of receptive but also of productive skills in the foreign language in which a text is presented for completion. If this is so, one can conclude that cloze is an integrative measure of foreign language skills.

The verification of the hypothesis was made on a population of more than 2,000 pupils in Estonia and the Ukraine. Altogether 15 kinds of receptive and productive language skills were differentiated. The performance was measured mostly by didactic (i.e., valid and reliable) tests with the exception of some skills evaluated by expert groups. The results in 15 skills showed the following correlations.

Table 5

Cloze as a measure of receptive and productive FL skills

	FL receptive skills:1)	corr.:	sign.:
CLOZE (English) (tested with more than 2,000 upper sec. school pupils)	reading comprehension	0.809	
	aural comprehension	0.843	
	FL productive skills:1)		
	spoken monologue	0.811	
	spoken dialogue	0.852	p (0.005)
	written dictation	0.680	
	written essay	0.836	
	integrated evaluation of FL proficiency2)	0.861	
CLOZE (German) (1,063 pupils of grade 7)	FL speech production:1)		
	dialogue	0.87	p (0.005)
	essay	0.93	

- 1) measured by valid and reliable didactic tests - e.g. characteristics of dialogue: number of syntactic patterns, variability of syntactic patterns, vocabulary richness (type/token ratio), number of replies to a partner, number of verbal stimulations to a partner
- 2) measured by expert evaluation

(Source: Rapoport et al. (1980).)

The method of verification has also been applied to the rate of deletion of words from the text (every 6th, 5th, 4th word or key words, i.e. those of the highest information value). It was found that cloze is most valid if every 6th word is deleted. Taken as a whole, the authors came to the following conclusions (Rapoport et al., 1980, pp. 113-116):

Cloze is an integrative test, i.e. it reliably reflects both receptive and productive language skills. Thus, cloze can be applied instead of a battery of special tests measuring particular kinds of FL proficiency. Cloze tests are suitable for mass application in schools, i.e. in large populations of testees as well as with smaller groups and individuals. The construction of cloze tests has to fulfil certain criteria: the texts have to reveal a logical coherence; every 6th word has to be deleted and altogether at least 50 words should be deleted; the scoring has to be done so that all contextually and grammatically correct words are accepted.11)

Thus, Estonian colleagues express their strong conviction that cloze procedure has advantages as a test of language skills in comparison with other tests. On the other hand they admit that cloze is not the only method which should be applied in FL testing - which is, again, in correspondence with the suggestions made by Enkvist and Kohonen (1978).

Notes

- 1 See, for example, the book by a leading Soviet pedagogue, Babanskij, Ju.K., 1982.
- 2 See especially Byčkovsky, P., 1982, and Skalkova, 1983.
- 3 Čajagi, L., 1978.
- 4 Normy ocenki znanij, umenij i navykov ucascichsja po inostrannym jazykam (Criteria for evaluating pupils' knowledge, skills and habits in foreign languages). Inostrannyje jazyki v škole, 1978, No. 1, p. 3-6.
- 5 Cf. van Ek, 1976.
- 6 Anglický jazyk - Učební osnovy pro I. a II. ročník gymnázia (English - Syllabus for grades 1 - 2 of gymnasium). Praha, SPN 1983.
- 7 See, for example, a special collection of papers Teorija i praktika jazykovych testov (Theory and application of language tests), ed. by R. Selg and I. Sotter (published by Research Institute of Pedagogy, Tallin, 1980). The book has been supplied with an English - German - Estonian - Russian dictionary of terms on language testing, pp. 203-229.
- 8 The FL aptitude of children living in the capital of Slovakia (Bratislava) and in towns was better than of those living in small (non-urban) communities. - Similar dependencies have been found in other investigations into psycholinguistic abilities and skills in

children, both in Czechoslovakia and in the USSR and other East European countries. This implies that the type of locality is probably a strongly relevant variable influencing the psycholinguistic profile of children living in a socialist society. I have discussed this problem in more detail (in relation to the validity of Bernstein's theory in socialist countries) in my review (of T. Erasmie's book "Language development and social influence") published in the International Journal of Psycholinguistics, 1978, No. 12, pp. 62-66.

- 9 The term "verbal intelligence" is not used in the terminology of Soviet researchers. Instead, the Russian term "umstvennyje sposobnosti" (i.e. intellectual abilities) is applied.
- 10 Enkvist, N.E. - Kohonen, V.
- 11 Much attention has been devoted to cloze scoring, in confrontation with the work of Bormuth and others, by J.A. Mikk, with respect, however, to measuring the difficulty of textbooks - see Mikk, J.A.: Optimizacija složnosti učebnogo teksta (Optimization of the complexity of textbooks). Moskva, Prosvesceniye 1981.

Bibliography

Babanski, Ju.K. Problemy povyšeniya effektivnosti pedagogičeskich issledovanij (Problems of increasing the efficiency of pedagogical investigations). Moskva, Pedagogika 1982.

Byčkovsky, P. Zaklady měření výsledku vyučy (Foundations of the measurement of instructional outcomes). Praha, ČVUT 1982.

Čajagi, L. Objektivizacia zistovania a hodnotenia vedomostí žiakov didaktickými testami (An objectivization of the assessment and evaluation of pupils' knowledge by didactic tests). Bratislava, Psychodiagnostické a didaktické testy 1983.

van Ek, J.A. The threshold level for modern language learning in schools. Council of Europe, 1976.

Enkvist, N.E. - Kohonen, V.: "Cloze testing: some theoretical and practical aspects." In: Kohonen, V. and Enkvist, N.E. (eds.), Text Linguistics, Cognitive Learning and Language Teaching. Turku, AFinLA, 1978, No. 22, pp. 181-206.

Gavora, P.: Rozvoj spoločnosti a modernizácia vyučovania cudzích jazykov (Development of society and innovations in teaching foreign languages). In: Rozvoj spoločnosti a výchova. Bratislava, Veda 1977, pp. 263-292.

Gavora, P. - Malikova, M.: Vyzkum cudzojazyčných schopností žiakov základnej školy (Investigating aptitude for foreign languages in basic school pupils). Cizi jazyky ve škole, 25, 1981-82, No. 4, pp. 143-153.

Gochlerner, M.M. - Rapoport, I.A. - Sotter, I.: O diagnostike sposobnosti k usvojeniju inostrannogo jazyka (Foreign language aptitude testing). In: Problemy obučeni-ja inostrannym jazykam. Tallin, NII pedagogiki Estonskoj SSR 1978, pp. 28-43.

Kokkosta, V. - Vyrk, E. - Murumets, S.: Issledovanie s pomosčju testa urovnja znanija angliškogo jazyka abiturientov Estonskoj SSR (Investigation of the English language proficiency of Estonian school-leavers by means of testing). In: Teorija i praktika jazykovych testov. Tallin, NII Pedagogiki Estonskoj SSR 1980, pp. 173-184.

Komkov, I.F.: Aktivnyj metod obučeniya angliškogo jazyka u učasčichsja IX klassa obsčeobrazovatelnoj školy (Motivation for learning English in pupils of the 9th grade of secondary school). In: Problemy obučeniya inostrannym jazykam. Tallin, NII pedagogiki Estonskoj SSR 1978, pp. 74-83.

Malikova, M.: Cudzojazyčne didaktické testy (Didactic tests for foreign languages). Bratislava, SPN 1978.

Rapoport, I.A. et al.: Testovaja metodika dopolnenija kak instrument psichologičeskich, psicholingvističeskich a metodičeskich eksperimentalnych issledovanij (Experimenting with cloze procedure in psychological, psycholinguistic and didactic research). In: Teorija i praktika jazykovych testov. Tallin, NII pedagogiki Estonskoj SSR 1980, pp. 87-125.

Rapoport, I.A. - Machlach, E.S.: K probleme izučeniya intellektualnych komponentov obučajemosti inostrannomu jazyku (An investigation into intellectual components of foreign language aptitude). In: Problemy obučeniya inostrannym jazykam. Tallin, NII pedagogiki Estonskoj SSR 1978, pp. 6-27.

Repka, R.: Sprava o vyučovaní anglického jazyka v 4. roč. gymnázia (Report on teaching English in the 4th grade of gymnasium). Cizi jazyky ve škole, 28, 1984-85, No. 1, pp. 1-8.

Rogova, G.V. - Nikitenko, Z.N.: O nekotorych pričinach sniženija interesa k prednetu "inostrannyj jazyk" u školnikov (On some causes of the decrease in pupils' interest

in the "foreign language" as a school subject). Inostrannyje jazyki v škole, 1982, No. 2, pp. 28-32.

Selg, R. - Tõevere, H.: Nekotoryje aspekty diagnostiki prigodnosti učebnyh tekstov (Some aspects of evaluating textbook texts from the point of view of their suitability for schools). In: Problemy obučeniya inostrannym jazykom. Tallin, NII pedagogiki Estonskoj SSR 1978, pp. 106-126.

Skalkova, J. a.o.: Uvod do metodologie a metod pedagogického výzkumu (An introduction to the methodology and methods of pedagogical investigation). Praha, SPN 1983.

Takač, J.: Provierka vedomostí z neslovanských jazyků u absolventů středních škol na VST v Kosciach - jazyk anglický (Tests in English set by the University of Technology of Kosice and assessing the mastering of non-Slavonic languages as shown by holders of secondary school certificates). Cizí jazyky ve škole, 26, 1982-83, No. 6, pp. 249-251.

Tomilov, V.A.: K voprosu ob unifikacii i specifikke lekaičeskogo minimuma anglijskogo jazyka dlja nacionalnyh škol (On the problem of unification and specification of a minimum English vocabulary for national schools). Inostrannyje jazyki v škole, 1978, No. 5, pp. 44-56.

Turek, J.: Učební výkony žáku SES v němčine (Learning achievements in German of the pupils of secondary economic school). Cizí jazyky ve škole, 24, 1980-81, No. 3, pp. 105-113, No. 4, pp. 151-158.

Turek, J.: Výzkum vyučování cizích jazyků v jazykových laboratořích a tradiční třídě (Research into foreign language teaching in language laboratories and in a traditional classroom). Cizí jazyky ve škole, 24, 1980-81, No. 8, pp. 339-348, No. 9, pp. 387-393.

FROM TESTING RESEARCH TO EDUCATIONAL POLICY:
INTRODUCING A NEW COMPREHENSIVE TEST OF ORAL PROFICIENCY

Elana Shohamy (Tel-Aviv University)
Thea Reyes (Bar Ilan University)
Yael Bejarano (Everyman's University)

1 Introduction

In Israel oral proficiency has been tested within the framework of the high-school leaving examination for twenty years. An oral test is administered by the Ministry of Education to all students at the end of the 12th grade. The common procedure is an informal, unstructured interview, conducted by a high school teacher who, by virtue of his experience in teaching, is called upon to administer the test. The topics of the conversation between the tester and the test-taker vary freely from everyday matters to more sophisticated subjects, without any advance planning. The scores assigned to the test-taker as a result of this 'test' are totally impressionistic, based on the tester's subjective judgement.

A number of deficiencies can be found in this procedure:

a) It represents a single aspect of oral performance, and it is therefore questionable whether it is a valid indication of students' overall oral proficiency. b) Since rater reliability is not measured, it is questionable whether the score obtained by the student is his/her true score, especially since the testers are not trained in either interviewing techniques or the rating of oral proficiency. c) The test has low variance and a narrow distribution of scores: literally nobody fails the test. It is clear that the test is not sensitive enough to tap the existing individual differences in oral proficiency.

Searching for an alternative to the existing system constitutes a problem, since among the large number of tests available (Madsen and Jones, 1981), hardly any have been sufficiently researched to allow their implementation on a nation-wide scale. The only oral test that has been researched extensively is the FSI (Foreign Service Institute) Oral Interview (Bachman and Palmer, 1981, Shohamy 1982, Hinofotis 1976, Clifford 1977). However, one of the major shortcomings of that test is that it does not encompass a wide enough variety of speech styles (Shohamy 1983); it is limited to questions asked by the tester and answers supplied by the test-taker. It is obvious that such a test cannot be a valid measure for assessing all aspects of oral proficiency. Recognizing this shortcoming, a number of U.S. government agencies have recently incorporated additional speech styles into the FSI Oral Interview procedure, such as reporting, conversing, and role playing (Lowe 1982). To date, however, there are no research data available on this newly revised test.

It is obvious that adding other speech styles to the Oral Interview can be justified from a linguistic point of view, since various speech styles utilize different linguistic and socio-linguistic features in oral discourse.

2 The purpose of the study

The aim of the research study presented in this paper is as follows:

- 1) To develop a number of experimental oral proficiency tests. The tests are intended to a) represent a broader range of oral speech styles which would presumably reflect the type of communicative language that students are likely to need in authentic situations, b) discriminate well among various levels of students' oral proficiency, c) have higher inter-rater reliability. These tests should also evoke positive attitudes on the part of the test-takers.
- 2) To examine the relationship between the scores obtained on the experimental tests and the scores on the existing oral proficiency test. It was assumed that if the psychometric quality of the experimental tests proves to be better, they may be considered a preferable alternative to the existing test.
- 3) To examine the extent to which the experimental tests are related to one another. In other words, whether any one of the tests can be a valid measure of overall oral proficiency, or whether there is a justification for the testing of the various speech styles by separate oral tests.

Four experimental tests were designed: Oral Interview, Role Play, Reporting Task and Group Discussion. The experimental tests were systematized in three ways: a) the tasks were constructed so as to discriminate among levels of performance; b) testers were trained to conduct and rate the tests, and c) scores were based on a systematic rating scale. These experimental tests were then compared to the existing oral test used in the national school leaving examination, for their psychometric properties.

This paper will report research results obtained on the development of the experimental tests, their relationship to the existing conventional test, and the inter-relationship among the experimental tests. These findings will provide the basis for recommendations regarding policy decisions.

Specifically, the following questions were asked:

1. Regarding the experimental tests: To what extent do they discriminate among various levels of oral proficiency and what is their inter-rater reliability? What are students' attitudes towards the tests?
2. Regarding the comparison between the experimental oral tests on the one hand, and the existing oral test on the other: Is there a difference between the means, standard deviation, distribution of scores, and correlation of the two types of tests?

3. Regarding the relationship among the four experimental tests: To what extent do students' scores on the different types or oral tests represent different speech styles related to one another and does the degree of relationship, by itself, suggest the need for different tests?

3 The design

The sample of the study consisted of one hundred and three twelfth grade students of four classes in a comprehensive high school north of Tel Aviv. These classes were randomly chosen out of seven classes in that school. All the students undertook all the experimental oral tests - the Oral Interview (OI), the Role Play (RPL), the Reporting Task (REP), and the Group Discussion (GD). The tests were administered on the campus of the Tel Aviv University; each test was administered independently and lasted for about ten minutes. In order to minimize the learning effect possibly created by the order of the tests, groups of students took the tests in various orders, so that total rotation was ensured. The OI, the RPL, and the REP were carried out on a one-to-one basis, while the GD was held in groups of four. Different testers were assigned to administer the tests; they were all experienced EFL teachers and were trained in administering and rating the tests. The rating of the oral performance was done on the spot, using a rating scale which was adapted from several existing rating scales, mostly from the DLI (Defense Language Institute).

It rated oral proficiency on a scale ranging from 4 to 10*, 10 being equivalent to native speaker's performance. The OI, RPL, and REP tests were audio-taped, and the GD was video-taped. This was done in order to provide the data for computing rater reliability, that is, comparing the scores assigned on the spot with scores given at a later date by independent, lay raters.

On completion of each of the four tests the students filled out a questionnaire which assessed their attitudes toward the four experimental tests.

Two weeks after the administration of the four experimental tests 77 of the 103 students were tested by the existing conventional test in their schools. It is on these 77 cases that the comparison between the experimental tests and the existing test was done.

3.1 The experimental tests

Four tests were used as the experimental oral test. The rationale for the selection of the experimental tests in this study was to tap a variety of speech styles which are likely to be needed in various oral communicative situations. It is realized that although these situations do not embrace the whole scope of oral

*The scale of 4 to 10 is the conventional scale regularly used in the Israeli School System.

performance, they provide a representative sample of useful oral interactions.

a) The Oral Interview (OI):

The rationale underlying this test was to guide the test-taker into a dialogue with the tester, which elicits answers to questions asked by the tester on different topics at different levels of proficiency. The test followed the model of the Foreign Service Institute Oral Interview (Lowe 1982). It was administered by trained testers and took about 10 minutes; it included the following four phases: 1) warm up; in this phase the test-taker was put at ease and the tester derived a preliminary indication of the test-taker's level of proficiency in speech and understanding. 2) level-check: during this phase the tester checked the functions and content which the test-taker could perform most accurately. 3) probing: in this phase the tester assessed the highest level at which the candidate could function accurately. 4) wind-up: in this phase the test-taker was returned to the level at which he could function most comfortably and thus was left with a feeling of accomplishment. The scoring of the test-taker's performance on the oral interview was done on the basis of the rating scale used for all the other tests.

b) The Role Play (RPL):

The rationale behind this test was to stimulate the test-taker to produce spontaneous speech-behavior within the limits of a pseudo-authentic situation on the one hand, and to elicit specific speech acts prompted by strictly limited situations on the other. The test was composed of a situation in which the test-taker had to play a role with the tester playing another, both partners in a dialogue. The test-taker was given a card on which he found the description of a situation and his expected role in it. The card was written in the test-taker's mother tongue, Hebrew, so that interference of reading comprehension in L2 should be avoided. The tester then engaged in the simulated conversation derived from the situation described on the card. The test lasted for about ten minutes, and a score was assigned by an assessor who was not involved in the RPL, on the basis of the rating scale used for all the other tests.

c) The Reporting Task (REP):

The rationale underlying this test was to stimulate the test-taker into a monologue in the foreign language based on authentic input in the mother tongue, Hebrew. The student was given an article in Hebrew which he was asked to read silently, and report its general content in his own words, in English. He was specifically asked not to translate the text, but rather to report freely, referring back to the text only if necessary. The test lasted about 10 minutes and was scored on the basis of the rating scale used to rate oral proficiency for all the other experimental tests.

d) The Group Discussion (GD):

The rationale underlying this test was to stimulate the test-takers

into a spontaneous discussion of a controversial issue, in which they could express views about topical matters, debate and argue over them, defend their opinions and try to persuade the other participants to accept them. Four students were asked to discuss a topical subject or issue controversial enough to lend itself to a lively discussion. Members of the group picked a card randomly from among 20 cards on the table. The card provided information regarding the topic of the discussion they were about to conduct and some clues for the discussion, such as guiding questions with some relevant lexical items. They were given a few minutes to read the card and plan the procedure of their discussion among themselves in their mother tongue, Hebrew, before they were to start the actual discussion in English. Then they discussed the topic in English (Reves 1982). The tester listened to the discussion without participating in it or interfering with it, and scored the performance of each of the four participant test-takers on the basis of the rating scale used for all the other tests in this study.

4 Results

4.1 Regarding the experimental tests:

a) Distribution of Scores:

Table 1 displays the distribution of scores obtained by the students on the four experimental tests. It shows that the distribution of scores for the four tests ranged along the whole scale of 4 to 10. In general, the distribution of scores on the experimental tests was similar.

Table 1

DISTRIBUTION OF SCORES ON THE EXPERIMENTAL TESTS

Score	OI		RPL		REP		GD		AVERAGE	
	f	%	f	%	f	%	f	%	f	%
4	2	1.9	3	2.9	1	1.0	9	8.7	4	3.62
5	9	8.7	13	12.6	10	9.7	9	8.7	10	9.93
6	27	26.2	31	30.1	30	29.1	27	26.2	29	27.9
7	38	36.9	33	32.0	34	33.0	40	38.8	36	34.9
8	11	10.7	10	9.7	11	10.7	7	6.8	10	9.48
9	10	9.7	8	7.8	12	11.7	5	4.9	9	8.63
10	6	5.8	5	4.9	5	4.9	6	5.8	5	5.40

N = 103

b) Inter-rater Reliability:

The extent of agreement of different raters regarding the oral proficiency scores was measured by correlating scores assigned by the tester on the spot (at the completion of the test) with a

second rating done on a stratified random sample of 25 selected tapes of each of the tests. It was carried out by two independent, untrained raters using the same rating scale as used for the first rating. The correlations were .91 for the OI, .81 for the REP, .76 for the RPL and .73 for the GD test, as displayed in Table 2. It is clear from these analyses that the reliability of the tests was relatively high, although it varied from one test to the other; the GD had the lowest and the OI the highest. We would assume that by assuring proper training of testers the inter-rater reliability could be even further improved.

Table 2
INTER-RATER RELIABILITY OF THE FOUR ORAL TESTS*

TEST	r
OI	.91
RPL	.76
REP	.81
GD	.73

*Based on 25 samples of each test

c) Attitude towards the four oral tests:

Attitude towards the four oral tests was assessed using a questionnaire that students filled out upon completion of each test. In the questionnaire students were asked to mark their agreement with statements which described the test as being difficult, interesting, instructive, etc. on a scale of 1 to 4, 4 meaning "strongly agree". Table 3 presents these results indicating that students favored the OI and the RPL: $\bar{x} = 3.05$ and $\bar{x} = 3.03$, respectively. The attitude towards the other two tests was less favorable, although still positive ($\bar{x} = 2.59$ for the Reporting and $\bar{x} = 2.74$ for the Group Discussion).

Table 3
STUDENTS' ATTITUDES TOWARDS THE FOUR ORAL TESTS

TEST	$\bar{X}$	S.D.	Order of Priority
OI	3.05	.44	(1)
RPL	3.03	.36	(2)
REP	2.59	.512	(4)
GD	2.74	.43	(3)

n = 103

4.2 Regarding the comparison of the four experimental tests with the existing test:

a) Differences in means and standard deviations:

Table 4 displays the means and the standard deviations of the four experimental tests and of the existing test for the 77 students who took both types of tests. The means of the four experimental tests were quite similar: they ranged from $\bar{x} = 6.00$ for the GD to $\bar{x} = 6.57$ for the REP. In contrast to the experimental tests, the mean of the conventional test was substantially higher, $\bar{x} = 7.79$, indicating the possibility that the scores may have been artificially inflated.

A substantial difference between the two types of tests was also found in the S.D. of the tests: while the S.D. of the experimental tests ranged from 1.32 to 1.93, the S.D. of the existing test was 1.03, indicating that the existing test is less sensitive to individual differences among students.

Table 4

MEANS AND STANDARD DEVIATIONS OF THE ORAL TESTS

Test	$\bar{X}$	S.D.
Oral Interview (OI)	6.49	1.39
Role Play (RLP)	6.17	1.51
Reporting (REP)	6.57	1.32
Group-Discussion (GD)	6.00	1.93
Existing Test	7.79	1.03

n = 77

A t-test analysis between the means of the existing test and each of the four tests indicated that these differences were statistically significantly different at the .001 level. Thus, a student may get a different score from the two types of tests.

b) Distribution of scores:

Table 5 displays the distribution of scores obtained by the students on the four experimental tests and on the existing test. It shows that while the distribution of scores for the four tests ranged along the whole scale of 4 to 10, the distribution on the existing test was very different: nobody received the scores of 4 and 5, while 13.55% of the students received these scores on the experimental tests. It is important to note that the scores of 4 and 5 are failing marks in the school system. It is worth noting what happened on the upper part of the scale. While about 56% of the students obtained scores of 8-9-10 on the existing test, only about 23% of the students received such scores on the experimental tests. The distribution of scores on the experimental tests did, in effect, resemble one another, but was still quite different on the existing test. These differences are even more noticeable in the polygraph displayed in Figure 1.

Table 5
COMPARISON OF DISTRIBUTION OF SCORES ON THE TWO
TYPES OF ORAL TESTS

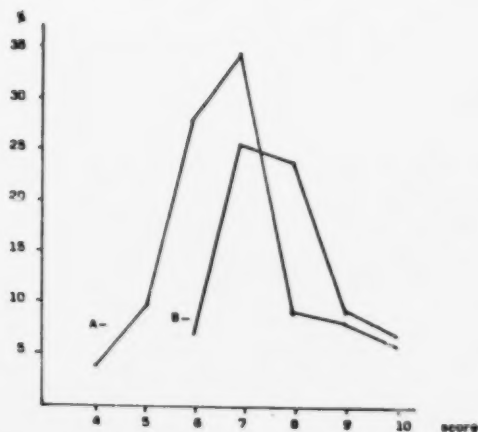
Score	Existing Test		Experimental Test (average of the four tests)	
	f	%	f	%
4	0	0	4	3.62
5	0	0	10	9.93
6	8	10.38	29	27.9
7	26	33.76	36	34.9
8	25	32.46	10	9.48
9	10	12.98	9	8.63
10	8	10.35	5	5.40

N = 77

N = 103

Figure 1

DISTRIBUTION OF SCORES ON THE TWO TYPES OF ORAL TESTS



A : AVERAGE OF FOUR EXPERIMENTAL TESTS
B : EXISTING TEST

c) Relationship between the Experimental and the Existing Test: Table 6 displays the r and the r^2 of the experimental tests and the existing oral test. It shows that the correlations between the OI experimental test and the existing test, which are based on the same speech style, interviewing, was .63, thus explaining about 40% of the variance. The mutual variance of the existing test and the other tests which tap other speech styles of oral proficiency is even lower, $r^2 = .23$ with the RPL, $r^2 = .29$ with the REP, and $r^2 = .20$ with the GD. It is, therefore, clear that in spite of the fact that all the tests were tapping oral proficiency, it is probably the different speech styles or the different methods of testing that brought the mutual difference drastically down.

Table 6
 r AND r^2 BETWEEN THE FOUR EXPERIMENTAL TESTS AND THE EXISTING TEST

Experimental Existing test	r	r^2
OI	.63	.40
RPL	.48	.23
REP	.54	.29
GD	.45	.20

4.3 Examining the Four Experimental Tests:

a) Correlations among the four oral tests: Table 7 displays the r and the r^2 of the four experimental tests. The correlations ranged from .60 to .70, that is 36-49% mutual variance among the tests. The GD had the lowest mutual variance with the other tests, while the OI had the highest.

Table 7
 r AND r^2 AMONG THE FOUR EXPERIMENTAL TESTS

TEST	RPL		REP		GD	
	r	r^2	r	r^2	r	r^2
OI	.69	.48	.70	.49	.60	.36
RPL			.67	.45	.64	.41
REP					.65	.42

5 Conclusions

Several conclusions can be drawn from these results. The first purpose was to develop and experiment with a number of oral proficiency tests which will represent a broad range of oral speech styles. These tests were expected to be accurate in terms of their psychometric properties and to evoke positive attitudes on the part of the test takers. The results obtained in this research indicate that the tests discriminated properly among the various levels of oral proficiency and had high inter-rater reliability (although it varied from one test to another). In addition, it can be claimed that by virtue of the fact that they tap a range of speech functions, the four tests, combined, have content validity. The test takers' attitudes to the tests proved to be favorable.

The second purpose of the research was to compare the experimental tests with the existing oral test. The existing test was found to have a restricted distribution of scores, limited variance, and the scores appeared to be inflated (practically nobody failed the test). In contrast the scores of the experimental tests covered the whole range of the rating scale, the variance was not restricted and the mean scores were more realistic. In other words, while meaningful information was obtained from the experimental tests, the existing test provided hardly any meaningful information. The correlations between the existing test and each one of the four experimental tests were not very high, even when the same speech style was tested. These findings allow us to conclude that, considering the above aspects, the experimental tests provide a preferable alternative to the existing oral test.

The third purpose of the research was to find out if and to what extent the relationship among the scores obtained on the four oral tests can provide evidence that the four speech styles are indeed different from a testing point of view. A low relationship could statistically justify the need for separate tests to tap overall oral proficiency. The mutual variance among the four experimental tests was found to be only moderate. This may provide evidence for the claim that the four tests indeed tap different constructs of oral proficiency.

6 Recommendations

The results of this research led us to recommend the introduction of a common measure of oral proficiency which consists of various speech styles. There is a strong educational advantage in expanding the number of speech functions in an oral test for high school students. It is most likely that the backwash effect created by a test encompassing multiple speech styles will give more emphasis to the instruction of these speech styles in the classroom. Clark (1980) calls for the development of such a test:

"...the successful development of a comprehensive 'common measure' proficiency test would make available a highly

refined and very powerful assessment instrument which could be used, in its own right, to evaluate communicative proficiency in situations for which the use of such an instrument would be considered both desirable and time - and cost - effective." (p. 25)

Regarding the structure of such a test which taps various language functions, there are two alternatives. One is a test composed of four isolated 'testing events', each conducted by a different tester (as for the experimental test in this study). Another alternative is a battery of tests, in which one tester tests the student in a number of functions in one 'testing event'. This latter version involves a separate rating for each speech function. An important question to investigate regarding these alternatives (isolated tests or a test-battery) is if the two types of oral test yield the same information for the test-taker.

7 From research to policy

The research reported in this paper demonstrates an example of the connection between research results and actual changes to be introduced into an educational system. Its main purpose was to evaluate an existing test and experiment with new tests, which, if found reliable and valid, would replace an existing testing procedure. The results of this research indeed indicate that there were good reasons to recommend the introduction of the new experimental tests.

The four-part oral proficiency test has indeed been accepted by the Ministry of Education to replace the existing oral exam. It will be introduced in 1986 as part of the nationwide Matriculation Examination taken by about 30,000 school-leaving high school students. It is an example of a situation in which the researcher can indeed be actively influential in creating a change in testing policy.

Bibliography

- Bachman, Lyle F. and Adrian S. Palmer 1981. "The Construct Validation of the FSI Oral Interview". Language Learning, 31:67-86.
- Clark, John L.D. 1980. "Toward a Common Measure of Speaking Proficiency". 15-26 in J.R. Frith (ed.). Measuring Spoken Language Proficiency. Washington, D.C.: Georgetown University Press.
- Clifford, Ray T. 1977. Reliability and Validity of Language Aspects Contributing to Oral Proficiency of Prospective Teachers of German: In J.L.D. Clark (ed.). Direct Testing of Speaking Proficiency: Theory and Application. Princeton: Educational Testing Service.
- Hinofotis, Frances B. 1976. An Investigation of the Concurrent Validity of Cloze Testing as a Measure of Overall Proficiency in English as a Second Language. Unpublished Ph.D. dissertation, Southern Illinois University.
- Lowe, Pardee Jr. 1980. Manual for LS Interview Workshops. Washington, D.C., CIA Language School. 1982 (revised) (mimeo).
- Madsen, Harold S. and Randall L. Jones 1981. "Classification of Oral Proficiency Tests", 15-30 in Adrian S. Palmer, Peter J.M. Groot and George A. Tropper (eds.). The Construct Validation of Tests of Communicative Competence. Washington, D.C. Teachers of English to Speakers of Other Languages.
- Reves, Thea 1982. "The Group Oral Examination: A Field Experiment", World Language English, Vol. 1, no. 4:259-262.
- Shohamy, Elana 1982. "Predicting Speaking Proficiency from Cloze Tests", Applied Linguistics, 3:161-171.
- Shohamy, Elana 1983. "The Stability of Oral Proficiency Assessment on the Oral Interview Testing Procedure", Language Learning, 33:3.
- Wilds, Claudia 1975. The Oral Interview Test. 29-44 in R.L. Jones and B. Spolsky (eds.). Testing Language Proficiency. Arlington, Virginia, Center for Applied Linguistics.

Bernd Voss (University of Bielefeld)

1 Introduction

The purpose of the present paper is to give an outline of the ways in which achievement in foreign languages is assessed within the context of the German school system.

It should be noted that in the above introductory sentence the term 'test' has been rather carefully avoided because - as will become fully apparent later - the assessment procedures used in the Federal Republic of Germany tend to be both less rigorous and more global than 'tests' in the usual understanding of the word (1). If we maintain that tests be objective, reliable, and valid, or that at least serious attempts be made to make them such, then with very few exceptions the procedures by which achievement in foreign languages is assessed within the German school system are not tests. In other words, there is no tradition of testing in the 'hard' sense of the word, in German schools.

This is partly encouraged by the fact that in Germany, as a Federal Republic, education is run not by a single national Department of Education, with resulting tendencies towards uniformity throughout the country, but - since education is constitutionally a matter of the individual states - by 11 independent Education Departments, each organising matters in somewhat different ways. In each of the parts of the country. In what follows we shall take the situation in the most populous state of North-Rhine Westphalia (NRW) as an example in the hope that this will at the same time mirror enough of the common ground that does exist between the approaches of the individual states, without, however, being able to claim correctness of detail for all the other states as well.

After a brief exposition of the organisatory framework of assessment in general, we shall follow the structure of the German secondary school system and deal with forms 5 - 10 and 11 - 13 as two different groups, before finally turning to the procedures used in the final (pre-university level) school-leaving examination, at the highest level of a student's school career. In each case, we shall present the official rules and regulations as issued by the Department of Education, and then try to explain in which way these tend to be implemented by teachers in the classroom, thus aiming at a survey of both the theory and current practice.

2 General Organisatory Framework of Assessment within the School System

Modern languages, as school subjects, operate under conditions

which are in no way different from those of other subjects. The appropriate assessment procedures, accordingly, need to conform to the general regulations that govern the ways in which the results of teaching and learning efforts are captured and measured within the system in general.

For North-Rhine Westphalia, this general framework is laid down in the Allgemeine Schulordnung (General School Regulations) of 1978, which stipulate, in § 21, that all assessment must refer to both knowledge and skills and needs to be based on all aspects of a pupil's work, in particular (i) on written pieces of work as elicited by a prescribed number of Klassenarbeiten (written classroom tests), which carry half the weight of the total assessment, and (ii) on what is called Sonstige Mitarbeit (other work), which accounts for the other half and is usually equated with the 'oral' aspect of classroom work although its definition includes participation and effort, and practical work as far as applicable, as well.

It is interesting to note that later (§ 22) a careful distinction is drawn between the written classroom tests as referred to above and limited written exercises as a further basis for assessment. These are not prescribed, may be marked, count for the 'other work' half with the value of a weightier oral contribution in class, but are not supposed to replace the assessment of oral skills. More details of this important distinction will be presented below (see section 3.2).

These arrangements imply (i) that assessment is conceived of as a continuous process rather than a once-off event, (ii) that in spite of explicit attempts to safeguard the assessment of oral work there is a noticeable bias towards writing as the basis for assessment, since writing-based procedures account for well over half the weight of the two assessment domains taken together.

While it is appreciated that procedures eliciting and evaluating work samples at several points over the educational year are more likely to yield reliable and valid results than those making the assessment dependant on a single, however massive, examination at the end of a course of instruction, language teachers find that the built-in bias towards writing-based assessment procedures makes it difficult to change the present practice of assessing the oral skills impressionistically, with no prescribed or customary method or format as a guideline, because it is in fact less important for the eventual result at the end of the school year and is considered to be the 'soft' part of the assessment, against the 'hard' writing-based parts. In other words, there is a strong tendency for the results in the written classroom tests - and possibly the marked written exercises - to dominate the evaluation of the achievement of a pupil at the end of the school year, with an independent assessment of oral skills either non-existent or only marginally modifying the assessment predetermined in this way.

3 Assessment of English (Forms 5 - 10)

3.1 Klassenarbeiten (Written Classroom Tests)

3.1.1 Organisatory Conditions

The importance of the written classroom tests for the assessment in general is suitably underlined by the range of rules and regulations that surround them.

As TABLE 1 shows, secondary education starts in the fifth form, which is the first year of English, and can be followed either in one of the school types of the traditional tripartite system, i.e. Hauptschule (Secondary Modern School), Realschule (Secondary Technical School) or Gymnasium (Grammar School), or in a Gesamtschule (Comprehensive School). Compulsory schooling ends after the 10th form. In all schools, English is compulsory. The table sets out how many lessons per week are held in English as a school subject, in each of the school types, how many of the prescribed written classroom tests are to be taken per educational year, and how long these tests are supposed to be (in periods per test).

Table 1
Written Classroom Tests in English (Forms 5 - 10)

year of English form	"Hauptschule" (Sec.Mod.School)				"Realschule" (Sec.Techn. School)			"Gymnasium" (Grammar School)			"Gesamtschule" (Compreh. School)	
	lessons per week	tests per ed. year	length of test in per.		lessons per week	tests per ed. year	length (periods)	lessons per week	tests per ed. year	length (periods)		
1	5	5	3	1	5	8	1	5	10	1	5	
2	6	5	6	1	5	8	1	5	10	1	5	
3	7	4	6	1	4	8	1	4	8	1	4	
4	8	4	6	1	4	8	1	4	8	1(-2)	4	
5	9	3	6	1	3	6	1-2	3	6	1-2	3	
6	10	3	6	1	3	6	1-2	3	6	(1-2)	3	(no data available)

It is interesting to note that although the number of English lessons per week is the same in all school types, in the more demanding schools more and longer written classroom tests tend to be required, a perhaps not quite convincing connection made by the system between frequency and length of measurement, on the one hand, and difficulty on the other.

All written classroom tests - as all other assessment procedures within the school system, with the possible exception of the Abitur (see section 6 below) - are set and marked by the teacher of the class. There is an explicit rule (Allgemeine Schulordnung § 22) obliging teachers to adjust the difficulty level of the individual test to the average standard of the class. This is implemented through the requirement that in the case of more than one third non-passes the test is to be repeated unless the headmaster of the school is prepared to sanction it, as an exception.

These arrangements are obviously designed to avoid any potential gap that might develop between the teaching and the testing of it: there is no problem for teachers to make sure that their pupils are tested on what they have been taught. It is equally obvious that the arrangements are intended to ensure that teachers do not get away with teaching without proper regard of what their pupils can learn: if more than a third of the pupils have not been able to follow to the degree of at least deserving the lowest pass mark for their efforts in the corresponding test, then the teaching was perhaps not as successful as it should have been, and an obligation to attempt to reach more members of the class seems reasonable. In this context, the 'exception' rule as outlined above apparently functions as a safeguard against possible misuse of the general 'one third' regulation by classes or groups wilfully pursuing a strategy of non-cooperation, or similar.

It is not difficult to sympathise with the pedagogical ethos that seems to have found expression in the regulations described so far. Unfortunately, these arrangements tend to have rather problematic side-effects in that they encourage a form of group-oriented norm-referencing - with each class functioning as an independent group - which amounts to a rather extreme version of what is of course in principle inherent in any system asking teachers to both set and mark their own tests, with few outside reference-points except the individual teacher's ability and conscience. It is temptingly easy for a teacher, when the results in a particular classroom test threaten to fall under the 'one third' rule, to simply raise the general level of the marking a little, rather than to go through the time-consuming process of setting and marking another test on the same field at a later stage: as long as the marking is internally consistent within the group, there is no immediate outside reason why this should not be done.

The consequence is that given marks are extremely difficult to compare across the system, as there are no established procedures of taking the different averages of the individual groups into account. Marks are therefore (more or less) useful indicators of the rank of

a learner in relation to his own group, but are practically uninterpretable as indicators of achievement across classes, schools, or regions: what is average in one group may be well above or below average in some other group of (formally) the same level, a natural consequence of accidental group composition and aggravated by systematic variations in the marking habits of teachers. These, incidentally, have reputations as 'hard' or 'soft' markers within their schools so that people (i.e. colleagues, pupils, parents) tend to appreciate a good mark from teacher X as equivalent to an average mark from teacher Y, etc.

There are two additional sets of organisatory rules connected with the written classroom tests which are of interest here (see Allgemeine Schulordnung § 22). To avoid surprise attacks, the dates of written classroom tests are to be announced in advance, they are to be marked, then returned to, and discussed with, the pupils (see also next section below), then given to the pupils to take home for their parents to see, and finally to be returned to the school. Further, they are to be spread evenly over the educational year, pupils are not to be given more than one on any single day, and there are to be no more than two in any week.

While both sets of rules make obvious pedagogical sense, the second one is rather important in contributing to the pressure to adjust the marking of a test, in the way outlined above, to avoid a repetition. Since pupils have to take written classroom tests not only in one subject, but in many others besides English, test dates are usually co-ordinated by the various teachers of a class and fixed well in advance, in order to meet the above organisatory requirements. Any later change threatens to overthrow the intricate balance of the various dates for the required number of tests in the different subjects, and the obvious result is that there is a strong interest in avoiding the repetition of a test altogether, as far as possible.

3.1.2 Formats

After the foregoing outline of the most important organisatory requirements connected with the prescribed written classroom tests it is now appropriate to look more closely into the nature of the formats used for this purpose.

Again, we can start from what is officially laid down, in this case in the Richtlinien (guidelines) of which there is a set for each of the subjects taught at school and which are issued by the Ministry of Education.

Taking the Richtlinien for the Grammar Schools as our main example (Kultusminister NRW ed. 1978b), the six years from the 5th to the 10th form are grouped into three two-year bands, with the following examples or recommendations for the written classroom tests:

- (1) forms 5/6: answer questions, form questions, describe picture, verbalise picture story, give simple explanations, change dialogue into narrative or vice versa, retell a text changing

the point of view (e.g. from 'He ...' to 'I ...'). A few techniques are explicitly not recommended, in particular dictations (neither of known texts, nor of slightly changed ones, nor of entirely unknown ones), the rendering of a text learnt by heart, the summary of a text (or part of it) read in class (cf. Kultusminister NRW ed. 1978b: 19).

(ii) forms 7/8: all of the above at a more demanding level, plus e.g. characterizations of people, objects, events; summaries (including structural analysis, main ideas, comments on sequence of events); analyses (at a fairly elementary level, at this stage) of stylistic/linguistic properties of given text types; changes from one text type (e.g. letter) to another (e.g. interview with letter writer); production tasks such as 'Describe the school day of a British pupil' (after treating the topic of school life in Britain in class over a period of time) (cf. op.cit. 37-38). There is an explicit note saying that integrative formats should definitely predominate, at this stage (cf. op.cit. 32).

(iii) forms 9/10: all of what is characteristic of the above group, only at a more demanding level and with particular emphasis on the analysis of fictional and non-fictional texts, e.g. work out the structure of the text, the pros and cons, plot, characters, historic background, comment on style/register etc. (cf. op.cit. 55). Again, explicit reference is made to the need of using integrative formats which alone can do justice to the complex objectives of the teaching at this level. This is considered to be more important than the loss in objectivity which this approach inevitably entails: "Die mit diesen Formen verbundene verminderte Objektivität der Leistungsbewertung muß mit in Kauf genommen werden." (op.cit. 56).

An attempt to see these guidelines in perspective reveals two interesting tendencies. The first is that written classroom work is clearly intended to set the model for the assessment formats employed to monitor the learning progress. In other words, the written classroom tests are to elicit work samples representative of the written activities performed in class, so that there is in fact an identity between exercise and assessment formats. Teachers tend to be rather in favour of such a scheme, to be sure for good pedagogical reasons.

The second tendency is less apparent in the letter than in the spirit of the guidelines, which implicitly favour a criterion-referenced approach - by outlining learning objectives and recommending assessment procedures asking for judgments as to whether and how far these objectives have been reached by the learners - in spite of, and in contrast to, the strong norm-referencing which is suggested by the external framework of the assessment system in general (cf. previous section). Whether these two contradictory approaches can be reconciled in a meaningful way, whether they cancel each other out to lose the advantages of them both, or whether this simply leads to a situation where anything goes and no obligation is felt towards any of

these principles, thus leaving the exact purpose of the assessment conveniently in the dark, is a question on which little information is available to date and which can only be raised rather than gone into here.

Practical realisations of the recommendations of the guidelines can be found in the Richtlinien themselves, which frequently draw attention to assessment formats when discussing exercises considered helpful in achieving the learning objectives. In addition, there are a number of useful publications available as help for the teacher (e.g. Schulz et al. 1984) which provide detailed discussions of concrete examples and often include marking and grading suggestions as well as handy bibliographies (2).

3.2 Tests (Marked Written Exercises)

The emphasis on integrative formats as described above does not mean that discrete item formats such as series of multiple choice, true-false, matching, fill-in or similar items, which are usually associated with objective testing, are not acceptable within the system of assessment at all. But they are considered mainly suitable for checking off some of the more isolated, mechanical aspects of language and are thus given a more limited rôle: tests consisting of tasks with these formats fall outside the regulations governing the prescribed written classroom tests, they are non-compulsory, and their results count towards the 'other work' area (cf. section 1 above; cf. also Kultusminister NRW ed. 1978b: 11).

Within the context of the whole system, this distinction, which is no doubt quite rational as such, unfortunately encourages two types of misunderstandings which contribute to making the 'other work' half of the assessment a rather problematic field.

The first is that in spite of explicit exhortations to the contrary (cf. Allgemeine Schulordnung § 22) discrete item tests, with their status of (non-compulsory) marked written exercises, are easily confused with a 'hard' way of measuring oral competence. In fact, even the Richtlinien themselves, in this case the ones for English in Comprehensive Schools, explicitly make this very assumption in a somewhat surprising statement suggesting that observing and assessing the oral command of the language can very well be carried out by informal written tests:

Es wird ... häufig notwendig sein, mündliche Fertigkeiten wie Sprechen aus Gründen der gesicherten und schnelleren Bewertbarkeit durch schriftliche Kontrollverfahren zu überprüfen.
(Kultusminister NRW ed. 1980: 142).

The second misunderstanding connected with discrete item tests within the present system is that they are in fact not quite so 'hard', i.e. objective, measures as one might think, because they are commonly implemented as quickly assembled lists of individual items, with little attention to the technique employed and the validity and reliability of the items included. In keeping with the tendencies of

the system in general, they are teacher-made and used individually for a particular group at a particular time. No pre-testing is done, nor are there any other steps taken to ensure that they can at least develop into anything more than isolated idiosyncratic snapshots of which language items the learners happen to 'know'.

Although these marked written exercises are commonly called Tests in the German school system, in contrast to the prescribed Klassenarbeiten (written classroom tests) described in the preceding sections, they are in fact rather remote from tests in the real sense of the word.

While teachers, understandably enough, often do not have the expertise, nor in particular usually the time, to improve the test qualities of these home-made 'tests' themselves, there is a somewhat surprising reluctance to use e.g. existing standardised tests, where after all this work has already been done for them. One may speculate that these are perhaps not sufficiently well known; it could also be that differences between various groups are so considerable that these tests are vaguely felt to be too remote from what is actually being taught in this particular classroom.

The alternative could be to develop the 'tests' into a network of criterion-referenced assessments, as another way of gaining a reference point outside the class for at least part of the total of the assessment procedures used. Again, it would be possible to draw on published tests of this nature, but again teachers appear reluctant to use them, possible for similar reasons as above.

The impression one gathers from an analysis of the Tests (in the curious understanding of the term given to it in the German school system) is that they are not an area of great interest or concern for teachers who are under enough pressure in other fields not to worry too much about this one. Considering the rather limited weight 'tests' have, as no more than non-compulsory parts of one half of the total assessment, this neglect is perhaps not totally surprising.

4 Correcting and Marking of Klassenarbeiten

What is accepted as more important, on the other hand, is the need for some kind of uniformity in the correcting and marking of the - more important - written classroom tests. In order to gain some insight into the problems connected with this field, it might be interesting to see which suggestions were made in a recent in-service training meeting (Landesinstitut für Schule und Weiterbildung November 1984), and to add some comments on them.

The problem of awarding an appropriate mark within the range of the marking scale prescribed by the system is only vaguely touched on by a reference to the learning and teaching objectives as laid down in the official guidelines (Richtlinien) already referred to previously: in how far these have been reached in a given written classroom test (Klassenarbeit) remains entirely a matter for the individual teacher to decide on and to convert into a mark. No guidelines are offered for

this most problematic aspect of any assessment. Guidelines are, however, considered more appropriate in connection with the correction and marking of mistakes.

All mistakes are to be marked and classified as to kind and gravity. The following categories are to be used and indicated by the appropriate abbreviation: Gr - Grammatik (grammar), T - Tempus (tense), Bz - Beziehung (concord), Sb - Satzbau (syntax), St - Stellung (word order), Pr - Präposition (preposition), W - Wort (lexical item), A - Ausdruck (collocation, expressions of more than one word), R - Rechtschreibung (orthography), Z - Zeichensetzung (punctuation). All of these count as one mistake, except for the last two which count only as 0.5 error unit. However, in dictations - which are considered acceptable only (if at all) at a rate of no more than one per half year in forms 5 and 6 - orthographic mistakes also count as one whole error unit. There is an explicit reminder that only those imperfections in the pupils' papers are to be counted as mistakes which can be shown (e.g. by dictionary, grammar book, reference book) to be definitely wrong: it is not enough for them to just 'sound wrong' to the teacher.

A number of additional specifications provide rulings for individual problems. There is a maximum of one mistake per word. A - mistakes, which by definition extend over more than one word, count as one single error unit, but other mistakes (e.g. orthography, grammar) count in addition, e.g. 'He locked after the key' would have one A and 0.5 R mistakes. Orthography counts as one mistake if the result is a different word, e.g. 'there holidays'. Similarly, if orthography is connected with grammar it counts as a (full) G - mistake, e.g. 'in it's nest'. Recurring mistakes are marked but counted only once; however, tense errors count each time unless there is an internally consistent tense shift; recurring elementary errors (e.g. in connection with the 3rd person singular -s) are treated as other recurring mistakes but if frequent are to be taken into consideration when determining the final mark.

After the correcting and marking, the written classroom tests are to be returned to the pupils in class. Teachers are expected to discuss the tests so that the pupils (i) understand why they have been given the mark they have received and (ii) can correct their own tests. The explicit idea is to help pupils to avoid making the same mistakes in future.

It can be assumed that similar suggestions are in principle in operation in many areas and countries. What is perhaps not quite satisfactory is that many of the error categories overlap considerably, so that the clarity suggested by the abbreviations is more apparent than real. Considering the orthography problems even of native speakers of English, the recategorisation of some orthographic mistakes as lexical (there - their) or grammatical mistakes (its - it's) is not entirely convincing. And in how far discussing the tests in class when returning them to the pupils contributes to helping them to avoid making the same mistakes in future is in fact rather a thorny question: there is little doubt that more systematic follow-up work than is pos-

sible within the one period usually devoted to it would be needed for this aspect of classroom work if the procedure is to be more than a ritual.

5 Assessment in English (Forms 11 - 13)

In the above outline of the ways in which achievement in English is measured within the German school system an attempt has been made to show in particular the general pattern of assessment during the six years of compulsory secondary schooling, with its emphasis on the written classroom tests, the problematic evaluation of the 'other work' area, and the curious rôle of the marked written exercise. When turning now to the final three years of instruction in English offered by the system to pupils attending Grammar Schools (and some Comprehensives) it will be sufficient to highlight the differences to this pattern, before finally looking in some detail at what is demanded in English in the Abitur, i.e. the final pre-university level school-leaving examination.

According to the Richtlinien for forms 11 - 13 (see Kultusminister NRW ed. 1981), the assessment of English at this level is intended to probe into three areas, (i) the command of the language as such, (ii) knowledge about e.g. language (→ linguistics), texts (→ literature), socio-cultural backgrounds (→ history, 'civilisation'), and (iii) mastery of methods (e.g. how to deal with an 'unseen' text, how to use a monolingual dictionary, a reference book etc.). These domains are not to be elicited and assessed independently of each other, but learners are to be given the opportunity of demonstrating their abilities in all three of them.

While the way in which this is envisaged will be sketched out below it is important to note at this stage that these demands constitute a definite shift and widening of the scope of assessment to include, above and beyond the foreign language as such, a strong content component (if we leave the 'method' aspect out of consideration for a moment), as a legitimate part of the total evaluation: in order to be successful in English it is now no longer enough to have an acceptable command of the language, but it is also important to be able to cope with the information retrieval and processing as it is required by the contents to be carried by the foreign language.

Since there is a considerable degree of overlap between what is demanded here and in other arts subjects (such as German as mother tongue, other foreign languages), in particular as far as methods, but also as far as some content areas are concerned (knowledge about language, text types, literary analysis etc.), high correlations between an individual learner's success in English and in these subjects can be expected. On the other hand, at a stage where a foreign language has been taught for over six years, language can only be expected to develop further in connection with contents that can be taken seriously and is a real challenge, so that there are many good reasons for broadening also the basis of assessment in this way.

5.1 Sonstige Mitarbeit ('Other Work') Domain

The basic pattern of assessment is still essentially the same as the one described above for the younger year groups, i.e. the assessment of achievement is based on (i) a prescribed number of teacher-made written classroom tests (now called Klausur), and (ii) the Sonstige Mitarbeit ('other work').

Generally, criterion-referencing is emphasized more strongly at this level, with the result that the 'one third' rule (see 3.1.1 above) is in fact somewhat relaxed (cf. Kultusminister NRW ed. 1981: 121-122; cf., however, 169).

There is a strong emphasis on the equal weight of these two assessment domains, underlined by the requirement to explicitly summarize, in a separate mark each, the results of the written classroom tests, on the one hand, and of the 'other work' domain, on the other, before drawing the two together into a single mark for the subject at the end of each half-year (cf. op.cit.: 122). The importance of the second domain of assessment is further emphasized by the fact that for some pupils in the final year, for whom the written classroom tests are under certain circumstances no longer compulsory, it may even constitute the only basis of assessment in the subject (cf. ibd.).

This makes it necessary for the Richtlinien to spell out in some detail how the domain of Sonstige Mitarbeit is to be interpreted (cf. op.cit.: 122-129). A basic distinction is drawn, for this purpose, between continuous assessment procedures on the basis of observation, and more formal assessments of individual pieces of work at individual points.

The first approach is considered to be particularly appropriate for the evaluation of the oral work done in class, where an assessment at one or two individual points only is felt to be too limited to capture a fair sample of a learner's contributions to oral classroom work. There is no doubt that the assessment of this field is intended to be the weightiest part of the 'other work' domain: there is an explicit note saying that - in conjunction with an assessment of the homework (see below) - it may well suffice as the basis for an assessment of the domain as a whole (cf. op.cit.: 124).

However, in spite of a detailed list of criteria that can be taken into consideration when evaluating oral performance at this level (cf. op.cit.: 149-150), teachers and pupils feel rather uneasy about the subjectivity of the impression marking that is practically unavoidable in this field. In a situation where teachers for good pedagogical reasons refuse to rush to their note-books to record a mark every time anyone says anything in class, and where they teach up to six different classes every morning, the only way for the conscientious teacher to arrive at a broad basis for the eventual assessment of the oral performance of his pupils is to systematically record marks from memory after school on the pupils taught that day. Under these circumstances, however, it is already sufficiently problematic to accurately recall who said what, to make a systematic application

of the criteria suggested practically illusory.

The second approach to the evaluation of the Sonstige Mitarbeit domain consists in the assessment of pieces of work elicited at individual points over the school year. Four types of work are recommended as particularly suitable, in a subject such as English, for this purpose (op.cit.: 124-129), viz. (i) homework, which at this level is considered to be complementary to classroom work and can e.g. also be of a preparatory nature, (ii) the minutes of a lesson, i.e. a somewhat detailed summary of the main points covered, (iii) marked written exercises (on which more below), and (iv) prepared talks on special, individually selected topics. Of these, only the marked written exercises lead to a direct comparison of all members of the same group; all the other types are usually interpreted as pieces of work which individual learners may be asked to do and hand in and/or read out in class, to be given a mark for individually.

This constellation tends to put the marked written exercises into a more prominent position. Consequently, a number of rules are in force to define their rôle more precisely and determine the organisatory framework in which they are to be used. As far as the latter is concerned, there are to be no more than two of them per half year. No more than two marked written exercises in different subjects may be taken by a pupil on any one day; when a pupil is due to take a written classroom test (in any subject) no marked written exercise is allowed for this pupil at all on the same day. Advance notice must be given, and the tasks set must allow completion within about 30 (and an absolute maximum of 45) minutes.

The marked written exercises are intended to provide insights into the ways in which pupils handle a particular limited though comprehensive question or problem connected with what is being taught at the moment. Series of short independent questions are explicitly ruled out, as are e.g. vocabulary tests (op.cit.: 127). Although in practice these latter stipulations are perhaps the regulations most frequently ignored in the whole context of assessment it is quite clear that the marked written exercises are meant to be short versions of the written classroom tests. Interpretations e.g. in terms of discrete item formats, which are still conceivable for the corresponding purpose in the lower forms (cf. 3.2 above), are not in accordance with what is intended, at this level.

5.2 Klausuren (Written Classroom Tests)

As pointed out above, the assessment of achievement is based, in equal parts, on the 'other work' domain, some details of which have been highlighted in the preceding paragraphs, and the prescribed written classroom tests, to which we shall turn our attention now.

The organisatory arrangements for this domain are in principle comparable to the ones for the lower forms (cf. 3.1 above). TABLE 2 shows that at this level English can be taken as an ordinary subject (Grundkurs) at three lessons per week, or as one of the two major

TABLE 2
Written Classroom Tests in English (Forms 11 - 13)

year in English	form	taken as			lessons per week	tests per half year	length of test (periods)
		ordinary subject (Grundkurs)		major subject (Leistungskurs)			
7	11.I	3	2	2-3	6	2	2-3
	11.II	3	2	2-3	6	2	2-4
8	12.I	3	2	2-3	6	2	3-5
	12.II	3	2	2-3	6	2	3-5
9	13.I	3	2	2-4	6	2	4-6
	13.II	3	1	3 hrs.	6	1	5 hrs.

subjects (Leistungskurs) required of every pupil in the final years, at 6 periods per week. The school year is divided into two halves (marked as I and II in the table), with in some cases slightly different regulations each. The final written classroom test in 13.II is given in hours instead of periods, as an approximation to the Abitur regulations (see section 6 below).

Learners are not to take more than one of these in any subject on any one day, and no more than a maximum of three per week. We have mentioned before that under certain circumstances some pupils are not required to take the written classroom tests in the final year at all: these pupils are not allowed to take them even if they wish - which they might want to do feeling happier with this kind of written work than having their assessment based entirely on the 'other work' domain - nor is it possible for them to be given a marked written exercise while the others do the Klausur. An interesting note stipulates that a written classroom test must have been handed back to, and discussed with, the pupils at least a day before the next one can be given (cf.

op.cit.: 169), as a - perhaps not entirely unnecessary - reminder to teachers of the pedagogical, in addition to the measuring, purpose of them.

As learners at this stage in their school careers are no longer taught in stable groups identical for most subjects but have in fact individual timetables regrouping them for practically all the subjects they take, a lot of organisatory dove-tailing - usually computerised these days - is required to meet the above conditions. Dates for the written classroom tests are worked out by the school administration, instead of the teachers, and are consequently felt to be prescribed from outside in a much stronger sense than in the lower forms. There is no doubt that pedagogical considerations, such as giving a test when the group is ready for it rather than two weeks before that stage or three weeks after starting on a new area already, plays no part in fixing the dates, on which teachers have practically no influence at all any more.

In keeping with the broadening of the assessment to include a strong content component as well, the evaluation of a written classroom test is based on an assessment (i) of the content, where errors would result from factual inaccuracies or logical inconsistencies etc., and (ii) of the language, where errors would result either from problems with formal aspects of the language (with correction rules similar to those outlined in section 4 above for the lower forms; cf. op.cit.: 168-169) or from problems with what is called 'text production' which is interpreted as the overall composition of the text produced by the pupil (cf. op.cit.: 167). It is recognised that language problems cannot always be attributed clearly to one or the other of the two language fields, but teachers are advised to take care that individual errors are counted only in one of them (op.cit.: 169). The overall mark awarded for a written classroom test is the result of a separate assessment of the content and the language, with the evaluation of the language as a rule the weightier factor (cf. op.cit.: 169).

The most problematic part of these regulations is perhaps the separation of the language part of the assessment into a formal aspect on the one hand and a style/composition aspect on the other. This is obviously intended to make it possible to strike a balance between these two - often non-correlating - aspects of a learner's written effort. However, the problems of operationalising the style/composition aspect into anything more than impression marking, or inspired guesswork, are considerable. Thus, it is perhaps not uncharacteristic of the situation that the overlap with the content area (when it comes to an assessment of the 'composition', i.e. the effectiveness of the presentation of the points made) which is after all parallel to the recognised overlap at the other end with aspects of the language form (in particular concerning lexical items), seems to have escaped the Richtlinien's notice. In practice, therefore, the evaluation of the style/composition aspect is usually considered to be the 'softest' part of the overall assessment which can be used to justify the mark which a test appears to deserve in the first place (3).

6 Abitur - Assessment in English

To round off our survey of the ways in which achievement in English is measured within the German school system, we can finally focus briefly on the arrangements made in connection with the pre-university level school-leaving examination (Abitur), for this purpose.

In accordance with the general Abitur regulations, the assessment in English, as it eventually appears on the certificate of the successful candidate in form of a mark, is in fact partly based on the results achieved within the final two years in the subject, and only in part on those obtained in the final examination itself, at a proportion of 2 : 1.

The final examination, on its part, can either take the form of a written test similar in type to the Klausuren described in the previous section, with an additional oral examination only if the results in the written part are markedly different from the average achieved throughout the final two years (or if the candidate hopes to improve his/her mark by taking one), or it can take the form of an oral examination alone in cases where candidates meet the required number of written tests in other subjects. While the complicated details of which candidates take which form of the examination are perhaps irrelevant here (for details see e.g. Kultusminister NRW ed. 1982), a short survey of aims, formats and evaluation procedures for both the written and the oral parts may be instructive (cf. Kultusminister ed. 1981: 169-235).

Similar to the Klausur regulations, the written examination is to give candidates the opportunity of demonstrating their command of the language, of particular fields of knowledge, and of adequate working methods.

To ensure some kind of conformity in the kinds of tasks set at this level, an agreement was reached in the Federal Republic in 1980 (cf. Sekretariat der Ständigen Konferenz der Kultusminister ed. 1981) to restrict the possible formats to the following four types:

- (i) The candidate is given a text (fiction) of about 700 words (major subject; time allowed: 5 hours) or 500 words (ordinary subject; time allowed: 3 hours) ($\pm$ 100 words) plus 4 - 6 questions specifying what is to be done with it, touching on content, form and an evaluation of the text and leading from comprehension to production.
- (ii) The candidate is given a non-fictional text, with tasks specified as under (i).
- (iii) The candidate is given a combined task 'Type A', which consists of
 - (a) a listening comprehension passage of 300 - 600 words (maximum of 4 minutes on tape) with 6 - 8 narrow questions eliciting

major content points of, and comments on, the text. The procedure is narrowly prescribed: (1) listen, (2) look at the questions, (3) listen again, (4) answer questions (time allowed: 30 minutes), (5) hand in. Only then is it possible to go on to the next part,

(b) a text-based task (fiction or non-fiction) similar in type to what is demanded under (i) or (ii) but 'shortened accordingly' (cf. Kultusminister ed. 1981: 177).

(iv) The candidate is given a combined task 'Type B', which consists of

(a) a picture, photo, cartoon or similar, accompanied by 3 - 4 questions eliciting a description and interpretation (maximum time allowed no more than one third of the total), and

(b) a text-based task similar to (iii .b).

In keeping with the tendencies of the whole system, it is the teacher of the candidates concerned who sets the tasks, which, however, in order to conform to some kind of common standard, need to be approved of by the regional Board of Education. The procedure is for teachers to prepare two packages of two different formats each and to hand these in to the appropriate Board of Education who select one of these packages (or reject asking for new suggestions) for presentation to the candidate who eventually chooses one of the two tasks included in it (4).

The criteria for the assessment of a candidate's work are in principle comparable to those for the Klausuren (cf. op.cit.: 184-187). Marking is done by the teachers who set the examination, with another teacher of the same school (usually appointed by the headmaster) acting as second examiner and the Board of Education exercising its power of inspection at its own discretion.

Of the four formats theoretically allowed, only the first two are in regular use, with the other two in practice occurring only as rare exceptions. This is of course not really surprising: the listening comprehension and 'viewing comprehension' sections of the combined tasks cannot build on any systematic preparation in previous assessment contexts and are therefore avoided as uncalculable risks; they are in addition surrounded by so many restrictive regulations (possibly originally designed to keep these two formats from developing into 'soft options') that they are technically unnecessarily difficult to handle; and the text-based formats are the most important parts in any of the four versions anyway, in the first two accounting for the whole, but in the last two still dominating the assessment as the weightiest factor by far, so that concentration on text-based formats appears rather more relevant - and sufficient - as preparation for the final examination (5).

The oral examination in English, as far as it is carried out (see above), also has a narrowly prescribed format. The aims are

similar to those of the written test, and the task is also text-based in that the candidate receives a text of about 300 words with 2 - 3 specifications eliciting aspects of content, form and comment from the candidate, who is allowed 30 minutes' preparation. The examination itself takes 20 - 30 minutes, the candidate is supposed to produce answers to the first two specifications as a continuous presentation with the third specification treated in form of an interview, i.e. by question and answer.

The oral examination is conducted by a small examination committee consisting of a chairman and three other members, one being the examining teacher, one acting as assessor, and one keeping the minutes. The examiner is required to prepare the tasks and pass them on to the examination committee at least two days in advance of the examination, for approval. It is the chairman's responsibility to ensure that all relevant regulations have been adhered to. If an inspector from the Board of Education is present, he would usually chair (some of) the orals.

It is interesting to note that in the oral examination listening comprehension tasks are explicitly excluded, if not entirely for the second part, then at least for the first (cf. op.cit.: 217). Another curious feature is the fact that minutes are to be kept which are to mirror rather precisely the most important aspects of the examination, including characteristic examples of the strengths and weaknesses of a candidate's language performance, but recordings are not permitted.

7 Conclusions

In the present paper an attempt has been made to provide an outline of the ways in which achievement in a foreign language such as English is assessed within the context of the German school system. After a brief exposition of the general organisatory framework of all assessment procedures at school, we have tried to describe the situation for the forms 5 - 10 and 11 - 13 of secondary education, as two different groups, before turning to the regulations that govern the conditions for the Abitur examination. It is hoped that the outline as presented here has provided a fair picture of the situation. Perhaps unavoidably, problematic aspects have been discussed in more detail than others. To redress the balance, attention should perhaps be drawn again to the great flexibility and freedom of a system which allows its teachers to tailor their assessment procedures to fit their teaching as closely as is conceivably possible, for most assessment purposes within the nine years in which English is taught at school. In addition, there seems to be an untapped potential for criterion-referenced assessment not sufficiently exploited so far.

On the other hand, there is little doubt that the extreme classroom-oriented norm-referencing demanded by the 'one third' rule encourages a situation in which outside criteria - i.e. questions such as 'How does a learner stand, and what can he do with his English, outside this particular classroom?' - unfortunately appear to be to-

tally irrelevant. This does, indeed, raise the questions of accountability in a rather serious sense.

As any other system that has grown and been modified in individual points over a longer period of time, the system described has its fair share of inconsistencies and problem areas, e.g. the assessment of the Sonstige Mitarbeit domain, the marked written exercises, the dominance of the text-based procedures with their backwash effect on the teaching, and many others. However, not all formats and procedures are problematic in themselves: if one were to pinpoint the single most effective measure to make the system more compatible with what professional language testers would have to demand, it would be a move to re-organise the marking. Rater training, multiple marking and independent marking would go a long way towards remedying a situation which in many ways in its present state of unsophistication would have to be called 'pre-scientific' in Spolsky's sense of the word (Spolsky 1975).

Notes

- 1 For a possible distinction between e.g. 'tests' and 'examinations' cf. British Council ed. 1976.
- 2 Examples of tests actually used for these year-groups were made available to the participants of the conference through the courtesy of Mrs Irmgard Voss.
- 3 A range of publications is available to teachers seeking suggestions, examples and guidelines on the problems of setting and marking a Klausur, e.g. Bliesener 1981a, 1982, Hurst et al. 1980, Multhaupt 1983 (for comprehensive bibliography). Examples of tests actually used recently were again made available to participants of the conference through the courtesy of Mrs Voss.
- 4 Again, examples were made available through courtesy of Mrs Voss.
- 5 For publications on the Abitur regulations (examples, discussions, criticism etc.) see relevant titles in the bibliography.

Bibliography

- Ahrens, R., and W. Hüllen. 1977. "Bemerkungen über die 'Einheitlichen Prüfungsanforderungen in der Abiturprüfung: Englisch'". Neusprachliche Mitteilungen: 25 - 29.
- Bliesener, U. 1981a. Klausur- und Abituraufgaben Englisch. München: Langenscheidt-Longman.
- 1981b. "Die neue Abituraufgabe Englisch gemäß den Einheitlichen Prüfungsanforderungen. Überlegungen zum sinnvollen Umgang mit

- der Textaufgabe". Praxis 28: 227 - 238.
- 1981c. "Arbeitsanweisungen im schriftlichen Abitur - Englisch". Praxis 28: 357 - 367.
 - 1982. Klausuren und Abiturarbeiten in Englisch. München: Langenscheidt-Longman.
- Bludau, M. 1976. "Normenbuch Englisch". Neusprachliche Mitteilungen: 68 - 70.
- The British Council (ed.). 1976. Examinations and Tests in English for Speakers of Other Languages. Information Guide No. 6.
- Hurst, N., Nissen, R., and U. Schulz. ²1980. Oberstufenarbeiten Englisch. Von der Textauswahl bis zur Bewertung. Dortmund: Lensing.
- Kornelius, J., and E. Otto (eds.). 1983. Anglistik. Richtlinien und Englischunterricht. Trier: Wissenschaftlicher Verlag Trier.
- Der Kultusminister des Landes Nordrhein-Westfalen (NRW) (ed.). 1978a. Allgemeine Schulordnung (ASchO) vom 8.11.1978. In Der Kultusminister (ed.) ²1979: 107 - 135.
- 1978b. Gymnasium Sekundarstufe I: Vorläufige Richtlinien Englisch. Köln: Greven.
 - ²1979. Handbuch Schulmitwirkung. Köln: Greven.
 - 1980. Gesamtschule: Richtlinien Englisch. Köln: Greven.
 - 1981. Gymnasiale Oberstufe: Richtlinien Englisch. Köln: Greven.
 - 1982. Die gymnasiale Oberstufe. Informationsschrift für Schüler. Bochum: Kamp.
- Multhaupt, U. 1983. "Nachgetragene und kommentierte Bibliographie zu den Richtlinien für die gymnasiale Oberstufe in Nordrhein-Westfalen, Englisch". In Kornelius and Otto (eds.). 1983: 200 - 225.
- Landesinstitut für Schule und Weiterbildung Nordrhein-Westfalen. 1984. "Fachübliche Grundsätze für die Korrektur und Bewertung von englischen Klassenarbeiten" mimeo.
- Pürschel, H. 1981. "Die Einschätzung und Bewertung mündlicher Leistungen durch deutsche Englischlehrer". Praxis 28: 368 - 375.
- Schulz, R., Eckardt, I., Semmler, M., and D. Williams. 1984. Klassenarbeiten im Englischunterricht der Klassen 7 - 10. Beispiel Realschule. Dortmund: Lensing.
- Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (ed.). 1981. Einheitliche Prüfungsanforderungen in der Abiturprüfung - Englisch. Beschluß

THE TEACHING OF TESTING:
RELIABILITY, VALIDITY, AND ALL THAT STUFF¹

Douglas K. Stevenson (University of Essen)

Two years ago - almost exactly to the day - I was in New York City at an ACTFL conference (that is, the American Council on the Teaching of Foreign Languages). Some friends and I were standing in a hallway outside a room, nervously awaiting our turn to speak. We'd been asked to be members of a panel, and the topic for our panel was practical applications of research in foreign language teaching. I'd been asked to take the language testing part; in other words, to deal with the question, "what does language testing research have to offer the foreign language teacher?" As I've said, we were a bit nervous. It was, after all, a very good question, and we were worried that someone in the audience might actually ask it.

While we were standing there, trying to think of some appropriate answers, two teachers came by. They were busily discussing that morning's program: which of several lectures they wanted to attend, which of the many panel discussions sounded interesting, or which demonstrations of new teaching techniques might be worth seeing. Needless to say, we were all listening in on their conversation, eavesdropping to hear their opinions. Then, after some discussion, one of the teachers pointed to a lecture in the program on a language testing topic. I've forgotten the exact title of the lecture, but it was something like "designing oral achievement tests for the classroom." "Now, that might be interesting," the teacher said. "Oh, no, we don't want to hear that, do we?" the other one replied. "They'll probably just be talking about reliability, validity and, you know, all that stuff!" Well, we thought, and laughed, that pretty much takes care of language testing, at least for them.

I'd like to talk to you this morning about an important area, an area that doesn't usually get the attention it deserves. This is the teaching of language testing to teachers. I'd like to first mention some difficulties we have, to sketch out the present situation. Then, in the second half of the paper, I'd like to talk about some content areas that need to be stressed more than they have been. Now, while I've thought about this topic, that mini-dialogue between two teachers, overheard two years ago in New York, has constantly been in my mind. It is there, itching at my memory, whenever the subject of language testing at school comes up. For a very good reason, I think, a reason we language testers tend to forget or, perhaps, would like to. Let me explain it this way.

Some of us here this morning might be described as language testing specialists (in addition, that is, to being teachers). Not necessarily experts, mind you. Rather, individuals - whether working primarily in the schools or universities, with testing centers or governmental agencies - who spend a great portion of their professional lives being actively and immediately concerned with the field of language testing in all of its many aspects.

Those of you who fit this description have probably by now spent several years just reading the language testing literature, those many articles, studies, research reports and proposals, dissertations and theses, conference proceedings, journals and books that are all related, directly or indirectly, to language testing. And, as you are only too well aware, this already vast literature keeps growing month by month and year by year. And, it's not just in Europe or even in North America that so much appears about language testing and assessment. For example, there's been the notably well-attended conferences in Singapore and in Hong Kong, the latest in a series. Moreover, what is being done in Japan, or in the Chinas, is also of considerable interest, especially to those of us concerned with the teaching and testing of International English. But whatever the specific area within the broad field of language testing, and from wherever it comes, be it Essen or Essex, Taiwan or Tampere, it will necessarily be of some interest to language testing specialists. To language testing specialists, language testing is, but of course, the major interest.

On the other hand, language testing specialists - like all specialists - are often so enamored of their subject, so fascinated by its complexities - that they overlook a simple thing. They tend to forget that their primary field of interest is not the ultimate academic delight for everyone else. And, I'm sorry to say, even in linguistics and language pedagogy there are those who feel, and have sometimes even stated, that the language teaching world does not revolve around language testing theory and practice.

We know that it should, of course. And, that's not just a feeling that derives from academic ethnocentrism, or so-called academic imperialism. After all, as Upshur and Oller and many other testers have pointed out, the assessment, measurement, and testing of language and language-related behavior - however conceived and carried out - is at the very center of all linguistics, language teaching, and language learning research.

Nonetheless, the fact that language testers often state this does not mean that others see it that way or, for that matter, are even disposed to consider it. Rather, they too as specialists are inclined to see their fields, their specialties, as constituting what language teachers really need. In fact, just about everyone in linguistics and language pedagogy is fighting for the attention of the classroom language teacher. Phoneticians insist that our future teachers must know a great deal about the etics and emics of their field. Sociolinguists, too, argue for more time with future teachers: it is obvious, isn't it, that sociolinguistic appropriateness and sensitivity bear at least as much weight in communicative situations as do accent and pronunciation? Even our colleagues who specialize in cultural studies will state their case to anyone willing to listen, and especially to the authorities who have required teachers to take courses in literature, linguistics, and didactics. Without cultural knowledge, they argue, these three areas produce only shallow pools of performance.

In this crowded, loud bazaar of language-based disciplines, the language tester has a very small voice, indeed, when it comes to attracting and holding the attention of the language teacher. What it

comes down to, I guess we could say, is the somewhat dismal fact that no matter how fascinating, and multifaceted, and fundamentally important language testing is to the language tester, it simply does not have, and never has had, the same inherent interest or felt-importance to the rest of the language teaching fraternity. And so, we might add, our sympathies go to them.

Now I do not find this situation, that language testing is not everyone's cup of tea, especially shocking. Nor would any of us, looking back over the past twenty years or so since Lado's Language Testing appeared (1961), expect the situation to change radically. It is simply the standard state of affairs, most everywhere, more or less, and if many of us are not content to live with it, we at least admit that it is there and should be changed. However, as I have suggested thus far, the topic does invite attention. There are several very basic reasons for this, each related to our central goal in language testing, which is to improve the quality of language assessment overall. Let me present the three most important of these, which are best considered as functioning assumptions.

First, you here today, that is, those of you who would not consider yourselves language testing specialists, do form a fairly unusual group. You are unusual, in a most positive sense. I hasten to add, because you are obviously willing to spend not a little of your time and schedules discussing language testing at school. There have been of course other such meeting and seminars. The IUS has sponsored some in cooperation with other groups. Furthermore, organizations such as International TESOL, or ACTFL, the British Council, and CITO have encouraged such gatherings. Yet all in all, it is still far from common to find groups of teachers voluntarily coming together to talk about testing. It is also still true that only a few of the hundreds of thousands of practicing teachers of modern languages throughout the world have had a chance to take any specific coursework in language testing. Usually, what is taken into the classroom is at most a few lectures. As I've discussed this situation in detail elsewhere (1983), I won't repeat it here. It can be concluded, however, that overall, language testing continues to be treated as if it were a nice, but not really necessary, subject.

Here we should remember that today's teachers are expected to be at home in a very wide range of disciplines and somehow - no one seems to spend much time wondering exactly how - they are expected to acquire all this knowledge within the little time usually available. It is, of course, unreasonable to expect someone to know much about what they have not been taught, or had a chance to learn. Simply to state, as is so often done, that "all good teachers" should know such and such, be it language testing or any other area, is to confuse an ideal syllabus with actual achievement. It all reminds me of a particularly humorless history professor we endured as undergraduates. When someone with enough courage finally asked him what would be included on his final examination, a fearsome four-hour affair, he gave a dry smile and said, "anything I haven't covered in class."

The first point, therefore, is that few teachers will be taught much about language testing during their teacher training period, and yet they are still somehow expected to know more than a little about it as classroom teachers. And certainly, whether or not language testing is a required subject in teacher training, and most often it is not, all of us would agree that the overall quality of language learning can only be raised by a thorough acquaintance with the principles, practice, and theory of language testing and assessment.

The second assumption is a conclusion that follows this first one. While language testing specialists are firmly convinced that what language testing has learned is very much worth transmitting, the opportunities to do so are few and far in between (with the notable, happy exception, of seminars such as this one). Now, someone might claim that what, with some dozen introductory language testing handbooks available in English alone, teachers who haven't had the chance to learn about testing during teacher training can simply go to their friendly, local bookstore, order one, and read it. Such a view is tenderly optimistic to the point of being naive.

All of us could make a long list of those textbooks in other areas that we will be reading sometime next week, no doubt. Or, perhaps next month. Or, more likely, considering all the work we have to do, sometime next summer. That is, of course, unless we have to move house, change diapers, get remarried, or read and grade all of those student's essays. Friends who are classroom teachers are usually so tired when they get home Saturday afternoon, that the last thing they want to do is to start reading a textbook on testing.

There is this Peter Pan view of language teaching that says, usually without thinking, that all teachers will somehow acquire expertise and knowledge that in some cases, even their professors have not had time to gain. To take this stance is hypocritical and unfair. Worse yet, it is unrealistic.

My second assumption, therefore, is that when it comes to the teaching of language testing to teachers, whether to those already in the schools or to those still hoping to get there, much of it, if not almost all of it, will have to be through non-traditional approaches and channels. Like this seminar, for example. Or through special, intensively advertised workshops. Or as adjunct meetings to regular pedagogy conferences. Or through series of introductory articles in popular, teachers' journals. But we do need to remember one thing. This is that the large gain in the number of professional, specialists' publications on language testing is not matched by a similar gain in the number of classroom teachers who have had specific and thorough instruction in language testing. The problem is to explain why they should spend much of their time learning about language testing when so often, the educational authorities do not see fit to make this central subject a required one.

This brings us to the third of the assumptions which are important when we talk about the teaching of testing. It is an irony, or at least a cause for some self-directed humor, that language testing specialists have spent so little of our time at conferences talking

about the teaching of language testing. Yet from the very beginning of language testing as a field, it was apparent that if language testing and assessment in the schools and universities was to improve, it had to be with the active cooperation of classroom teachers. It is worth remembering here that Lado's Language Testing: The Construction and Use of Foreign Language Tests had a subtitle, A Teacher's Book. And, as far as I'm aware, almost all language testing books that have followed, have followed Lado's example and have been written primarily for, and directed to, the teacher.

It is also appropriate here to note that the Interuniversitäre Sprachtestgruppe, the IUS, was founded to pursue two goals. First, there was a need, at least in our educational venue, to establish an informal, frequently-meeting forum where specialists in language testing could talk to other specialists and thereby be aided and abetted in their efforts. Secondly, there was an obvious need to provide a friendly, informal setting in which non-specialists could acquire general to specific knowledge about language testing, and thereby be aided and abetted in their efforts, as well. We'd noticed that at most conferences - even those that had themes connected with teaching in the schools - there were few teachers and fewer students in attendance. Specialists talking to specialists, professors to professors, was the rule. We had also noticed that teachers who did not really know much about testing were understandably unwilling to stand up at formal meetings and ask questions that might be considered naive. Finally, we were well aware that the inherent interest in language testing would not be sufficient to have many paths beaten to our international doors. Therefore for all these reasons, the IUS would need to be a welcoming, rather than a closing or closed, organization.

What should be well-understood, especially as it is of fundamental importance to the teaching of testing, is that this second goal did not just spring from altruistic concerns or, God forbid, from any warped sense of public relations. Rather, the two goals are intertwined, at least in educational systems in which teachers have a say in what is taught and how what is taught is tested. No matter how good our tests are, no matter how refined, how attentive to current theory, psychometric principles, and pedagogical good house-keeping, they really don't stand much of a chance without the active understanding and support of the classroom teacher.

We're not just concerned with newly developed or better testing techniques and measures. As has often been said, and is still so obviously true, the worst testing practices that are so commonly encountered do not result from any intent on the part of those who use them to be unfair to pupils. Rather, they are generally traditional in that reliability is of little interest. And authority is usually more than sufficient to take care of any embarrassing questions about consistency or content. It is therefore all rather simple, it might seem. All one must do is to bring the two groups together, the testers and the teachers.

Or so it might seem. About three or four weeks ago, I just happened to be present during one of those discussions among col-

leagues that the outsider might see as a heated debate, or even as an argument. The topic of translation testing, that is, the grading and scoring of translation examinations, somehow came up that morning between classes and coffee. I was just an observer; if not neutral, at least silent, for I wanted to hear what was said about translation testing. For some time I'd been trying to find out more about this subject, still so important in our university system. I'd been checking with specialists in the area, people who had been trained both as translators and as testers. These specialists had been in general agreement that injecting reliability into translation grading would encounter the usual problems of what constitutes validity-based variance, and what error, and so on. But, following the usual analytic or holistic approaches to grading long used with essays, establishing a high degree of reliability posed no real problems, at least from the language tester's point of view. In short, the argument did have my full attention, as these colleagues were much more sure of themselves and their positions than the specialists I'd talked with had been.

What it all came down to was that those present that morning did not believe that testing had anything worthwhile to say about translation grading. One colleague stated in no uncertain terms that all such testing approaches he'd read about were essentially nonsense (although that was not the exact word he used). He had taught for years and years. He had read more translations than he cared to remember. He had graded translation examinations until his eyes glowed red, like some neon rabbit. And he knew for a fact that the grading of translation was subjective in fact and in practice, that no consistency could be reached, and that nothing, let alone a few correlational coefficients would ever change that. There was no use even talking about seeing new ways to score translations, and that was that, period, full-stop, punkt!

One factor we should keep in mind, therefore, is that language testing is frequently an emotional issue. A great many people already have very strong opinions on what is good and bad, possible and impossible. Moreover, those who are strongest in their opinions are often those who actually know the least about language testing.

As much as we would counter these negative feelings, we can't overlook the very very positive benefits of securing teacher support and understanding. These can equally be given example, and two come readily to mind. First is the incredibly wide-acceptance in North America in foreign-language teaching circles of what many of us know best as the FSI Oral Interview procedure. The second example is the graded objectives movement in Great Britain which we've had the pleasure of hearing more about during this seminar. What is striking about both movements, although an ocean apart, is not, as one might expect, the actual tests or testing approaches. Both have their antecedents in the established testing literature, so what is being done in either area is not really new. What is so striking is that each has spread and has been so seemingly successful because each has been actively supported by classroom teachers. This did not come about by chance.

Specialists working with the oral interview, for example, spent several years presenting this approach through just about any channel available, both professional and popular. Groups such as ACTFL, the Interagency Language Roundtable, TESOL, ETS, and individuals working in schools and universities came together to explain the oral interview and to answer questions. Teachers were gone to, talked with, and listened to when they questioned this problem or that detail. The results have been simply astounding.

We should note that this publicizing has not in any way been like the selling of a soap, a new deodorant, or the latest toothpaste. Those most responsible for presenting the oral interview have also been most responsible in pointing out its limitations, constraints and, above all, the great effort and training needed to make it effective. This nationwide movement in the US, therefore, is far from being a mindless surge. We need to mention this last factor, for there is always the specialist's fear that in making anything better known to the non-specialist, it will necessarily be simplified, or cheapened in some way.

Experienced language teachers, however, are already skeptical about any teaching approach that claims to be the one best way of teaching anything or anyone. Most of us, for instance, are already past counting on one hand the number of times a so-called new approach or method has been proclaimed as the quickest and surest way to language learning Nirvana. Most of us have also discovered from experience in the classroom what others have established through research: no single teaching method fits everyone teaching or everyone taught.

There are of course in language testing - as in any field - those who would sell a testing product that cures all academic illnesses in record time, that can be administered anytime by anyone for any reason to everyone. But language testing specialists are in a good position to compare the one to the other - the teaching cure-all to the testing cure-all - and thereby lessen the appeal of both.

This third assumption, therefore suggests that probably the single best way to improve the quality of testing in schools is not necessarily by introducing new tests or new testing techniques. Rather, it is to make available, to make better known, the critical tools already available and used in language testing.

I'd like to pause here for a moment, and sum up some of the points I've tried to make thus far in talking about the teaching of testing. First, language testing specialists are in agreement that their field - at its broadest - is at the heart of linguistic, language learning, and language teaching research and methodology. More specifically, they are convinced that the intelligent application of language testing approaches, techniques, and methods to the classroom can do a great deal to improve the overall quality of language learning and teaching. However, at the same time, this viewpoint is not shared for the most part where it most counts. A relatively small number of in-service teachers have been specifically trained in

language testing. A relatively small number of teacher training programs require specific coursework in language testing and assessment. The language teaching community might pay lip service to the importance of language testing and here and there offer active support. Nonetheless, overall, few language testers have the opportunity to teach more than a little bit about testing to more than a few future teachers.

Secondly, there is already a large amount of literature about language testing set at the introductory level, and specifically directed at the modern language teacher working in a second, or foreign language classroom. For the most part, this literature is readily available. In some cases, free materials have been made available by various non-profit organizations. However, at the same time, because language testing is seldom given the same five-star ranking accorded other disciplines in teacher training, it cannot simply be assumed that the practicing teacher will see fit to purchase and read these materials. Certainly, other disciplines compete for his or her time. And certainly, teachers always have less time than might appear to those who do not teach.

Thirdly, and this should be stated as bluntly as possible, language testers cannot go it alone. We cannot assume that what we say and write will somehow naturally be of interest to those outside our specialty, just because we think it should be. Specialists might all be in agreement that a certain new test is damned fine, or another one potentially as dangerous as the Black Plague in its effects on pedagogy or the pupil. But this means very little in either the success of the one, or, sadly, the eradication of the other. This is no doubt a hard lesson for us to learn. But the fact that in educational systems marked by great overall sophistication and strong tradition, markedly unfair examinations still persist - and sometimes are even dominant - should forcibly remind us that cooperative action is not nice, but ultimately necessary. There are, as well, strong attitudes, both negative and positive towards types of testing and testing approaches already in existence. As a result, language testing must be prepared to go out and deal with what it considers to be unfounded prejudices, misinformation, or even, clichés and stereotypes about testing that are of the common garden variety. Every good test must have 100 questions, right? Of course not.

Basing these assumptions upon previous literature written on the teaching of testing, research into attitudes towards tests and testing, and conversations with language testing specialists internationally, this is the situation, generally speaking, that we must deal with when we teach, or would teach, about language testing. More or less, of course, for each educational system, each teacher training program, has its own tendencies and trends. A few do require specific training in testing (Israel is a good example). Others simply list it as one of those abilities teachers should have. Still others would not even list any language testing book on the standard "important or required" reading lists distributed to all incoming students. Macbeth, yes. Oller, no.

Someone once said that none of us likes to take a good, long, realistic look at anything, be it the state of the contemporary world, our personal finances, our children's rooms, or even the teaching of testing. I do think it would be wrong, however, to be anything but guardedly optimistic, or even out and out positive, about what can be done, and in some systems, is being done in the teaching of testing despite this overall situation that I've briefly described. Again, you are a very good sign of the times, as it were, a most positive indication that language testing is being viewed more and more as a very worthwhile area for the classroom teacher to spend time in and effort on. Someone is buying those language testing volumes. No one really expects Madsen's new (1983) volume on testing, published by Oxford University Press, for example, to meet the best seller levels of a novel like Fear of Flying. But it has appeared, another one of many good ones, and will be read by many.

Also operating in favor of language testing are two general but continuing educational trends. First, and perhaps most important, there is that concept known as accountability. Simply stated, more pressure, whether coming from governments, tax payers, parents or pupils, is being placed on those who teach to be responsible and public about their successes and failures. In some systems, for instance, teachers are now required by law to demonstrate their own proficiency in the language they teach. In others, teachers are paid to the extent that they can demonstrate that their pupils have achieved what they should have achieved.

The second general trend affecting a stronger interest in testing has made itself more and more felt since the end of the Second World War. The spread of education at all levels to groups not previously encompassed has also led to a general breakdown of that authority that so characterized traditional testing approaches in the schools. Again, simply stated, fewer parents, or pupils, or students are willing to accept the teacher's word as being the final word of God and government.

Over twenty years ago now, an acquaintance of mine teaching at a university in Michigan was taken to court by a student who thought that she deserved a "B" on an essay examination she'd written, when he'd given her only a "D." Experts were called in to testify about reliability. The essay was read by groups of specialists to determine inter- and intra-rater reliability ranges. He won the court case, finally, but it didn't do much for his nerves. Now, you might be thinking, that sort of nonsense can't happen here, wherever "here" might be. But it did happen there, then, and had, as you can well imagine, a very strong effect on the standards and care of essay grading. Suddenly, testing was very relevant, and how one can grade essays reliably a topic of more than casual or intellectual interest. Students not only take tests; they take them personally.

These two general trends might be seen as constituting negative sources of motivation to learn more about testing. True, but I've met quite a few language testers who, having tried to explain the importance of reliability and such to sceptical audiences, feel that those who have ignored such questions so intimately related to ethics

and fairness deserve what they get. There's nothing personal, mind you, it's just that they feel that certain practices have continued too long to be tolerated any longer. If they're forced to choose between the teacher's comfort and the pupil's welfare, most testers feel they have no choice.

What various legal authorities, above all in North America, are now saying about classroom testing is a most interesting subject. But here, it just needs mentioning as one of the several factors that is creating more of an interest in the literature of language testing. The point is that despite the handicaps we must realistically face in the teaching of testing, much can be done, and much must be done.

In the time left, I would therefore like to turn to what we might call the content material and matter. Now, you are very sensitive, as the result of lectures during this seminar, about content validity. Aware of this, I will state at the outset, that the suggestions I have for various areas that should be more stressed in teaching testing do not claim any universal, or specific validity for any given educational system. Importantly, they are offered as fundamental bases and considerations in the teaching of testing, considerations that have been too often slighted in more usual approaches to the topic. I've chosen four areas that I think we should emphasize more than we have.

As content in the teaching of language testing, I do believe and, over time, have become more firmly convinced, that the concepts of reliability and validity should be first and foremost in any list of what must be understood. "Understood" is the key word here, for a good feel for the concepts and their interrelationship is, in the long run, much more important than a quick review and a few exercises in how one, for instance, calculates a test-retest statistic. Just about everything else in language testing does intersect at some point with these two, and time spent insuring that students do have a good feel for them is well spent, indeed. Both need to be given their full due as the central concepts and not just treated in passing, or placed as afterthoughts.

There is another reason for according reliability and validity this central and primary position. And not understanding this is in fact why the topic of reliability and validity is so often greeted with a "yeah, yeah, we've already heard of that" sort of reaction. What separates the traditional testing approaches from the modern, the language tester from the metrically-naïve layman, is not specific types of tests or testing approaches. It is an appreciation of the nature of, and the need for, reliability. This is encompassed in that measurement dictum that "a test can be reliable without being valid, but to be valid, it must be reliable."

When this concept is grasped, the inter-relationship appreciated, a sort of "aha!" reaction often sets in, and the basic importance of language testing becomes clearer. It is an ice-breaker, if you wish, as a topic. It would not be too much of an exaggeration to say that teachers are converted, at this point, and never quite look at testing

in the same way again. One thing that Alan Davies was getting at in the first issue of Language Testing (1984:112) can be furthered here. It is very easy, especially with computers just about everywhere, to introduce the topic of reliability, and soon have people turning out data. But we should not confuse the ability to perform various statistical manipulations with an thorough understanding of the concepts. It certainly helps, but it's not the same.

To explain reliability so that students feel it, takes a good teacher and a good bag of metaphors and analogies. For example, I've noticed that in Harrison's most recent introductory language testing volume (1983:10f.), he uses an elastic measuring tape as a metaphor. "There would be little point in trying a measure people's waists with a piece of elastic," he notes. And, "the tape measure should be placed so that it goes around people's waists, not round their hips or the waistband of the trousers they are wearing."

Reliability and validity do have the reputation of being esoteric areas, and those who think that the contemplation of construct validity theory is good for the soul probably don't have many sympathizers. But the general concepts can be introduced in a relatively short period of time, and the effect of their being understood is a dominant one on whatever else follows.

A couple years ago, for instance, Uli Raatz took a fresh group of students at a workshop in Bielefeld through the concept, and pointed out what reliability had to say about our grading system in German schools. The atmosphere was electric when the two were brought together. "That can't be true," was of course one reaction to be expected. The other one, sort of "well I'll be damned" (and quite a few pupils obviously have been) was another. But at this point, how one can express and determine reliability made more sense, for there then existed clear reasons for doing so. For validity, it is perhaps enough here to repeat Spolsky's well-known phrase (1975:153), "the central problem of foreign-language testing, as of all testing is validity."

If we take reliability and validity and their interrelationship together, then it can be stated that an understanding among teachers if their significance can do more good, more quickly, than anything else to improve the quality of testing in the schools. Everything else, all that other stuff, more or less follows these principles, with the possible exception, depending upon definitions, of the "why" of testing. It is most unfortunate, I think, that reliability and validity have come only to be associated among many in language areas with only certain types of testing, as if they were things that Lado and his followers thought up to make our lives more difficult. On the other hand, if we look at standard introductory testing volumes, it is as if language testers, having accepted these principles, assume somehow that everyone else has, as well. They are often presented as givens, instead of as concepts that must be continually presented, defended, and explained to each new generation of teachers. In any teaching of testing, they must have first place, always.

The second area of language testing that we need to emphasize more than we have goes in the opposite direction, that is, from the heart of the forest to the trees. This is what we could call the over-all educational and social context, and functioning of testing. The type thing Spolsky did with his three trends in language testing article would fit here, of course. But what I'm thinking of is a larger historical perspective, the type thing more often done in educational and psychological measurement courses, where the field of human measurement is surveyed and exemplified. If we wish, we could call this a perspective on the "why" of language testing, today and yesterday, in this place, and that.

We would want to include as much comparative information as is possible, that is, what systems other than ours are doing and why. Those lectures we've heard here about testing in different educational systems, for instance, are of great value to those of us who will, when all is said, not have to work within them. They offer us new ways of looking at our own cases. They challenge our own assumptions. Furthermore, and this is perhaps most important, they tend to knock down that stubborn view - the one I exemplified earlier with reference to translation testing - that something simply can't be done, or isn't worth doing.

For years, we were told that oral interviews simply couldn't be carried out on a continent-wide basis and still have any comparative value. But this is what appears to be happening in the US today. In another direction, the work on self-assessment measures done by Tibor von Elek and his colleagues often comes as a pleasant surprise to those who think a testing device must always assume the examinee is somehow lying or, at best, untrustworthy. As a third example, there is the TOEFL. What it is and what it does just doesn't make sense to anyone who doesn't know the background of educational needs in North America that made it necessary. Too often it has been judged simply as a test in isolation, and not as part of a larger measurement procedure in a larger educational context.

Yet we should not just begin with our own time or just within our own systems. It's too easy to be parochial in both, to assume what we do is somehow natural, or the only way. A standard article given in introductory educational and psychological measurement courses is entitled "A test-dominated society: China 1115 B.C. - 1905 A.D." Americans, who might otherwise see their own nation-wide, testing programs as both recent and singular, usually benefit greatly by viewing a society that had similar programs some two-and-a-half thousand years before Columbus got lost on his first trip to India. We would also want students to see and, if possible, take as many different types of tests as possible. "Do unto yourself, before you do unto others," is a good testing rule. We would always want to survey what is included in the broad range of interests we call language testing. Just to show how much is available is important. It also tends to prevent us from going out and constructing what we think are new measures, before we've taken the time to see what is already available.

The third area that I would suggest needs to be given much more attention is related to this second one. This is what is commonly

termed language learning research. Here, we would want to point out that classroom language testing is only one small area where language tests strongly influence language teaching and learning. Almost every single research study we can think of has, somewhere, usually at its heart, a language test of some type that serves as the fulcrum upon which teaching recommendations will be based, and new approaches claimed and defended.

What appears in the standard, international language teaching journals is fertile field for such material. Many of our students, for example, now find themselves taking courses about second and foreign language acquisition. There is that famous, or to many language testers, infamous Bilingual Syntax Measure that sits right in the middle of many interlanguage theories and research studies. It's validity is a central question. Attitudinal and motivational research is fast becoming of greater interest in many systems. We have Hauptschule students who are now required to take another year of English, for instance, and what might still interest them, is a good question indeed. Adult education, bilingualism, and those languages losing their fair share of students to English have also accelerated interest in this area. Again, right in the middle are all those tests and testing techniques. Are they reliable? Are they valid? Are they appropriate?

We could go on, certainly, but it should be clear that language testing is a most useful applied tool for anyone, whether or not he or she will ever actually construct a test battery. It's interesting to look back over recent issues of popular journals, those supposedly directed at teachers such as the TESOL Quarterly. More and more, the pages are filled with statistics and references to various measurement concepts. Not a few of our students, probably most, simply skip across those pages, those correlational tables and such, just as if they were written in a foreign language that they didn't know. Here language testing can perform a badly needed interpretive service. It is wrong that many teachers should be unable to understand, let alone judge, so much of what is presented today. The misuse of testing, tests, and measurement concepts is a very big problem and here we can do much, rather easily, to restrict it, as well.

There is one other aspect to this emphasis that is less applied, but just as important, I think. This is the issue of complexity. An increasing methodological rigor provided by testing in all areas of language studies has led to an increasing awareness of just how complex it all is. Ironically, if someone is looking for a simple, straight-forward answer - and who isn't? - there is a natural tendency to shy away from this complexity. Often, as a result, both research and tests which promise the most are also those which are least sophisticated in measurement terms. They are, therefore, the most appealing to those with not much background in testing. We need to deal with this problem head on, and why, for instance, a test that is not pretested, or does not report say, reliability statistics or even SEMs is probably not a good one, is usually a good place to start.

This brings us to the fourth and final of those areas that I

think we need to stress more within that "all of it" possible universe of content. It gets to the heart of why so many of us are so persistent in talking about the teaching of testing. It brings together, as well, our concern with the quality of testing and our emphasis on reliability and validity. This is the entire question sets of ethics and accountability and fairness.

All language testers must operate, as they know, within a set of what might be called functional ideals. There is, for example, no such thing as a perfectly reliable test. Knowing this, we still manipulate the concept in theory and practice as if there were. Likewise, there is no such thing as a perfectly valid test for all purposes, pupils and conditions. Yet we also manipulate this ultimate constraint as if there could be. We do the best we can, knowing all the time that from the very beginning, the nature of testing and the extent of our knowledge about what is to be tested will compromise it. We know from the outset that it will be far less than ideal.

One reason language testers have always been seriously concerned with the teaching of testing is that perhaps more than others, they know what a tricky business it all is, even at its best. With a constant eye on the educational and social consequences of testing, they must, as a matter of professional and personal conscience, continually point out the distance between best practice and ideal. A good knowledge of reliability and validity, therefore, also imposes upon one an awareness of the responsibility that connects the why, what and how of testing with decisions based upon it. And certainly, knowing the iffiness of the entire structure we are much much better off than if we didn't. We might not be much happier, or more content, but we are more responsible for what we teach and, more importantly, to those we teach.

The direct connection between what are often disparagingly termed "only" psychometric considerations, and the effects testing has on what is taught, who teaches, and who is taught, is also one reason that language testing is a rather exciting place to be. By its very nature, it tends to challenge tradition, authority, and not least, simple smugness. Those coming to testing for the first time often expect a somewhat reactionary activity, and are therefore often surprised just what revolutionary potential is always there. Approaching the field of language testing through discussions of fairness, and accountability, is often not a bad way to start, for it sparks off a greater interest in what it's all about.

To conclude here, we should state of course, that everything else in that "all of it" of language testing is also important and, given time and world enough, we would also teach. However, the four areas which I've suggested we might want to give more emphasis to are seen as basic structures upon which the rest can be built. Even if a classroom teacher is never called upon to construct a test battery, even if a teacher never carries out his or her own empirical research, they are highly useful. Overall, I hope I've been able to strengthen our mutual view that the quality of testing stands to gain as much from the teaching of testing as it does simply through more testing.

And finally, that "reliability, and validity, and you know, all that stuff," is of central importance to any teacher, any where, any time.

Notes

- 1 This was the summary paper for the Heinola teachers' meeting, and the oral-presentation style has been preserved here.

Bibliography

- Davies, A. 1984. Review of Oller, J.W., Jr. Issues in language testing research. Language Testing 1/1: 111-115.
- Harrison, A. 1983. A language testing handbook. London: Macmillan Press.
- Madsen, H. 1983. Techniques in testing. New York: Oxford University Press.
- Spolsky, B. 1975. "Language testing - the problem of validation." In Palmer, L. and Spolsky, B. (eds). 1975. Papers on language testing 1967-1974. Washington, D.C.: Teachers of English to Speakers of Other Languages. 146-153.
- Stevenson, D.K. 1983. "Foreign language testing: all of the above." In James, C.J. (ed). 1983. Practical applications of research in foreign language teaching. Lincolnwood, IL: National Textbook Company. 153-203.

ÄIDINKIELEN TAIDON MITTAAMINEN LUKION PÄÄTTÖVAIHEESSA

Anneli Kauppinen (Ylioppilastutkintolautakunta)

TIIVISTELMÄ

Äidinkielen taitoa testataan Suomessa lukion päättövaiheessa ainoastaan esseetyyppisellä kirjoitelmalla. Koetyyppi ei ole sata-vuotisen historiansa aikana kokenut sellaisia suuria muutoksia kuin vieraiden kielten kokeet. 1920-luvulle asti äidinkielen koe oli samalla myös reaalikoe. Aiheet ovat tietysti väljentyneet ja monipuolistuneet, mutta edelleenkin annetaan yleensä vain muutaman sanan aiheita, joiden pohjalta essee on laadittava. Lukion uuden oppimääräsuunnitelman ja päättökokeen välillä on ristiriita, sillä lukion kirjoittamisen opetus tähtää monipuolisuuteen.

Ylioppilastutkintolautakunnassa on kokeiltu muutamia uudenlaisia tehtävätyyppejä, kuten referaattia, mielipidetekstiä ja kirjallisuusanalyysiä. Esseemuotoinen koe aletaan joka tapauksessa säilyttää mahdollisten uusien koetyyppien rinnalla.

Ylioppilastutkinto eli lukion päättökoe on toinen opetussuunnitelma. Se ohjaa opetuksen suuntaan ja tavoitteenasetteluja konkreettisemmin kuin varsinainen opetussuunnitelma. Siksi olisikin tärkeää, että nämä kaksi olisivat tavoitteiltaan samansuuntaiset. Parhaassa tapauksessa päättökoe konkreettistaa opetussuunnitelmaa.

Tätä nykyä lukion äidinkielen oppimääräsuunnitelma ja ylioppilastutkinnon äidinkielen koe ovat vasta etsimässä toisiaan. Ylioppilasma ei ole satavuotisen historiansa aikana kokenut sellaisia selviä muutoksia kuin muiden kielten päättökokeet, vaikka käsitys hyvästä kielitaidosta ja samalla opetuksen tavoitteista on äidinkielessäkin muuttunut.

Uuden lukion oppimääräsuunnitelmaa mukaillen voidaan sanoa, että lukion päättävän olisi kyettävä käyttämään kieltään seuraaviin tarkoituksiin:

- tiedon vastaanottamiseen ja välittämiseen,
- ajattelun selkeyttämiseen,
- tunteiden ja elämysten vastaanottoon ja välittämiseen,

- esteettisen mielihyvän saamiseen ja tuottamiseen,
- yhteydenpitoon.

Oppimääräsuunnitelma pitää äidinkieltä ensi sijassa taitoaineena. Siksi kokeenkin pitäisi mitata taitoja mahdollisimman monipuolisesti. On kuitenkin syitä, joiden vuoksi koe ei voi olla kattava. Niitä ovat ensinnäkin käytännölliset ja taloudelliset rajoitukset, toiseksi tradition voima ja kolmanneksi tutkimustiedon rajallisuus.

Käytännöllisistä ja taloudellisista syistä koetta on tietysti aina pelkistettävä ja yksinkertaistettava, niin ettei kielitaidon koke kirjo voi tulla siinä huomioon otetuksi. Niinpä on tuskin mahdollista mitata valtakunnallisessa kokeessa esimerkiksi kokeilaiden suullista ilmaisutaitoa. Käytännön syyt voivat myös johtaa sellaisiin koejärjestelyihin kuin ovat vieraiden kielten monivalintatehtävät. On luotu systeemi, jonka periaatteet on lukiossa opetettava eräänlaisena aputietona mutta jolla ei ole paljon tekemistä kielitaidon kanssa.

Tradition voimasta koe on voinut pysyä samantyyppisenä pitkäänkin, ja tässä onkin yksi syy äidinkielen kokeen nykyisiin ongelmiin. Koska aine on alun perin ollut sekä äidinkielen että reaalien koe, sen jotkut tehtävänannot mittaavat yhtä paljon tietomäärää kuin kielenkäyttötaitoa. Tämä on johtanut siihen, että jotkut opettajat katsovat parhaaksi opettaa kirjoittamista faktojen kautta, niin ettei voi sanoa, onko kysymys tietojen vai taitojen opettamisesta.

Aineesta sanotaan, että se antaa kokeilaille vapauden persoonallisiin ratkaisuihin, ajattelun hiomiseen ja oman näkemyksen välittämiseen. Nämä perinteisen aineen hyvät puolet on reilusti tunnustettava. Aineen väljyydessä ja vapaudessa on kumminkin sekä tämän tehtävätyypin voima että heikkous. Uskallan epäillä, että aine ei suinkaan anna käytännössä mahdollisuuksia sellaiseen vapauteen kuin mielellään mainostetaan. Ylioppilasaineissa on harvoin sellaista reipasta mielipiteellisyyttä kuin nuorilta ihmisiltä odottaisi. Onko meillä väärä kuva nuorista vai leikkaako siivet

opettajan tai sensorin pelko? Käsitys kielikuvien tuoreudesta ei ole samanlainen kokelailla ja aineiden tarkastelijalla. Parempi siis välttää kuvakieltä, joka kenties onkin "kulunutta" tai tyyliin so-pimatonta. Mistäpä nuori tietäisi? Mistäpä kukaan niin tarkkaan tietää? Mitä kaunokirjallisempaa on aineen tyyli, sitä hyllyväm-mällä suolla opettajat ja sensorit ovat arvosanaa määritellesään. Ansaitseeko yltiöromanttinen, paikalleen pysähtynyt luonnonkuvaus paremman arvosanan kuin täsmällinen ja johdonmukainen mutta tyyliiltään tavanomainen asia-aine? Opettaja pystyy suuresti ohjaa-maan oppilaidensa tyylivalintoja, kunhan hän saa tiedon sensorien määrittelemistä vapauden rajoista.

Nykyisellään aine-nimike kattaa moninaisen kirjoittamisen alueen. Aiheisto on laajentunut koulun oppiaineista nykypolitiik-kaan, joukkoviestintään ja oppilaan harrastus- ja kokemuspiiriin liittyviin asioihin. Samoin on tehtävänäntöjen muoto monipuolistu-nut: pyydetään kirjoittamaan reportaasi lehteen, kehoitetaan omien tunteiden pohdintaan tai runojen erittelyyn. Jotkut aiheet edellyttä-vät tiukan asiallista käsittelytapaa, jotkut taas sallivat vaikkapa pakinan.

Ketään ei tietenkään haltaa, että aineeksi kutsutaan mitä moninaisimpia kirjoittamisen lajeja. Ongelmallista on kuitenkin se, että kaikkiin sovelletaan samoja arvosteluperusteita. Vaikka tehtä-vänannossa onkin jokin tekstilajiin viittaava ohje, sitä ei tarvitse käytännössä noudattaa - tarkkaan noudatettavia ovat sen sijaan joka tapauksessa oikeakielisyyshormit.

Koska kielentutkimus on perinteisesti kiinnittänyt päähuomion virkkeiden, lauseiden ja sanojen muotoa koskeviin seikkoihin, ne myös saavat erityishuomion aineita arvosteltaessa. Tarkoitukseni ei suinkaan ole asettaa korrektin yleiskielen vaatimuksia kyseenalai-seksi, jos väitän, että helposti tunnistettavia oikeakielisyyserik-keitä saatetaan tarkkailla suurennuslasilla ja samalla jättää muut, jopa olennaisemmat tekstin laatuun vaikuttavat seikat huo-miota. Kokeen tavoitteiden määrittely ja suoritusten evaluointi on

paljolti sidoksissa siihen, millaista tietoa kielestä on olemassa, miten hyvä ja huono kieli määritellään ja nimenomaan: kyetäänkö se kyllin tarkasti määrittelemään.

Vaikka aineen sanotaan olevan tekstimuodoltaan vapaa, käytännössä noudatetaan tavallisimmin edellisten vuosien ylioppilaskokeissa esitettyjen laudaturaineiden tavallisia tekstinormeja. Tällä alueella olisi paljon sellaista, mikä tehtävänlaistajien tulisi saada kiteytetyksi selviksi tavoitteiksi. Ongelma on se, että tekstimuodoista on saatavissa niukalti tietoa ja sekin ristiriitaista. Pyrkimyksiä aineen muodon väljentämiseen on jo sentään näkyvissä. Ylioppilaskokeita 1984 -kokeen toimittajat ovat lähteneet luokittelemaan esimerkkiaineita sen mukaan, mitä tekstilajia ne lähinnä edustavat. Vaikka tekstilajien ominaispiirteet ovat epäselvät, monimuotoisten ainetekstien esittely väljentää ilahduttavasti aiempaa kaavamaista käsitystä hyvän aineen muodosta. Seuraava askel onkin se, että tehtävänannoissa joko selvästi osoitetaan että muodon valinta on vapaa tai sitten annetaan selvä tekstilajin määrittely, jonka noudattamista tarkkaillaan selkeillä kriteereillä. Ylioppilastutkintolautakunta onkin jo nimennyt työryhmän laatimaan uudenlaisia, täsmällisempiä arvosteluohjeita erilaisia aineratkaisuja varten.

Tekstilingvistiikan merkityksestä kirjoittamisen tavoitteiden määrittelyssä, opetuksessa ja arvioinnissa on monenlaisia käsityksiä. Yksi melko yleinen luulo lienee se, että tekstimuotojen ja tekstien ominaisuuksien määrittely kaavoittaisi ja kahlitsisi kirjoittamista. Minun käsitykseni on nykyisin aivan päinvastainen: jos tieto tekstien ominaisuuksista on vähäistä ja yksipuolista, kaavoittamisen vaara on suurempi kuin silloin, kun tietoa on paljon. Tekstintutkimuksen tehtävänä ei nimittäin mielestäni ole luoda normeja, siis esimerkiksi "jalostaa" tekstejä joidenkin mallien mukaisiksi - normit luo kieliyhteisö. Tekstintutkimuksen tehtävänä on auttaa näkemään tekstin ominaisuudet selvemmin ja pelkän intuitiivisen tuntuman asemasta myös määrittelemään ne. Näin saadaan apua sekä kirjoittamisen tavoitteiden määrittelyyn että arviointi-

perusteiden täsmentämiseen. Sellaiset deskriptiiviset määrittelyt kuin "töksähtävä", "sekava" ja "kehittynyttä ilmaisu" ovat kovin köykäisiä tekstin ominaisuuksien mittareiksi.

Aineen rinnalle on ylioppilastutkintolautakunnassa viime vuosina suunniteltu muitakin tehtävätyyppejä. Kokeilussa on tähän mennessä ollut referointitehtäviä, kirjallisuusanalyysseja ja mielipidet tehtäviä. Kaikille näille on ominaista, että vaadittavat tekstit ovat ainetta lyhyempiä. Kahdella ensiksi mainitulla pyritään mittaamaan myös tekstin lukemisen taitoa.

Nykyisessä äidinkielen päättökokeessa ei juuri tähän asti ole käytetty hyväksi mahdollisuutta laajentaa taitojen mittaamista kirjoittamistaidosta lukemistaidon arvioimiseen. Kun oppimääräsuunnitelman ja koko nykyaikaisen koulutuksen keskeisiä tavoitteita on tiedon vastaanottamisen opettaminen, on merkillistä, että asia on jäänyt äidinkielen kokeessa lähes huomiotta. Vähäistä tämänsuuntaista kehitystä on tosin ollut nähtävissä, sillä vuodesta 1974 lähtien on yhden aiheen ohessa ollut lisämateriaalia: kaava-kuva, taulukko, piirros, asiateksti tai kaunokirjallinen teksti. Lisämateriaali on tarjottu kirjoittamisen tueksi; sen ymmärtämistä ei yleensä ole suoranaisesti testattu, paitsi kun on ollut kysymys kaunokirjallisen tekstin analyysistä. Jotkut tehtävänannot ovat olleet sellaisia, että oppilas ei ole voinut tietää, onko aineen perustuttava annettuun materiaaliin ja missä määrin.

Jos äidinkielen kokeessa halutaan mitata asiatekstin ymmärtämistä, yksi mahdollisuus on monivalintatehtävä ja toinen - paljon parempi - referointitehtävä. Laajahkoa asiatekstin selostamistehtävää onkin ehditty jo ylioppilastutkintolautakunnassa kokeilla ja tutkia. Kävi ilmi, että aiempaa tutkimustietoa ei referaatista juuri ollut saatavilla. Siksi työ vaatii melkoisesti aikaa.

Referaatista ollaan päättökokeen tehtävätyypinä monta mieltä. Äidinkielen kokeen kehittämistyöryhmä valitsi sen yhdeksi tehtävähedokkaaksi, niin että se voisi olla valinnaistehtävänä aineen rinnalla. Vaikka referaatti tuntuu monesta kovin yksinkertaiselta ja vähän ajattelukykyä vaativalta tehtävätyypiltä, kokeilu osoitti,

että suomalaiset lukiolaiset eivät ole siihen harjaantuneet, eivät edes kaikki ne, joiden ainekirjoitusarvosana oli kiitettävä. On kysymys eräänlaisesta kommunikatiivisen kielenkäytön perustaidosta: kyvystä ymmärtää teksti, erottaa siitä olennainen ja selostaa se selkeästi ja lyhyesti uudelle lukijalle. Tämä kyky näyttää olevan toinen kuin se, mitä vaaditaan aineen kirjoittajalta - arvosanat eivät nimittäin korreloi keskenään.

On aihetta pohtia, miksi magnan tai laudaturin tasoinen aineenkirjoittaja laatii ymmärrettävästä asiatekstistä huonon referaatin. Voidaan kysyä, kumpi vaatii enemmän ajattelutaitoa ja kommunikointitaitoa. Oikeaa vastausta ei tietenkään ole, koska aine on kovin monimuotoinen ja tekstilajiltaan määrittelemätön. Useat aineet kyllä osoittavat kirjoittajansa ajattelukykyä, mutta ylioppilaskirjoituksissakin voi saada magnan sujuvasta tekstistä, joka ei välttämättä edellytä johdonmukaisen ajattelun taitoa. Referaatti saattaa onnistua ainetta paremmin sellaisilta oppilailta, jotka kykenevät kirjoittamaan lyhyesti, napakasti ja selkeästi lehtikirjoittamisen tapaan - ainekirjoituksessahan sanamäärän suuruus ja siitä monesti johtuva pitkäjäykkisyys katsotaan usein ansioksi. Koulussa on pakko tätä nykyä keskittyä pitkien tekstien harjoitteluun, koska on olemassa tietoja siitä, että hyvät ylioppilaslaineet ovat yleensä pitkiä ja heikot lyhyitä. On syytä kysyä, onko pitkän ja jähkailevan kirjoittamisen muoto oikea ja ainoa ihanne nyky-yhteiskunnassa. Eikö sen rinnalle pitäisi jo tuoda lyhyen, nopeasti kommunikoivan kirjoittamisen ideaali?

Tutkijan kannalta sellaisen selvärajaisen tehtävän kuin referaatin kokeilu on hyödyllinen, koska se pakottaa selvittämään asiatekstin perusominaisuudet. Tärkeäksi nousee esimerkiksi kysymys siitä, mitkä tekstin ominaisuudet vaikuttavat sen referoittavuuteen. Esimerkiksi tavallisen kertomuksen tai kuvauksen referoiminen ei ole vaikeaa. Sen sijaan tekstit, joissa on erittelevää ainesta tai mielipiteellisyyttä, vaativat referoijalta enemmän. Erittelevät eli ekspositoriset tekstit vaativat mutkikkaiden merkityssuhteiden ilmaisemiseksi täsmällistä käsitteiden määrittelyä ja

harkittua sidoskeinojen käyttöä. Lukijan on tuo tekstuaalinen aparaatti hallittava ja vielä kyettävä siirtämään alkuperäiset merkityssuhteet referaattiinsa. Mielipiteellistä tekstiä referoitaessa lukijan on selvitettävä ensin itselleen ja sitten uudelle lukijalle, mikä lähtötekstissä on kirjoittajan mielipidettä ja mikä muiden näkemystä. Mielipidetekstin eräänlainen dialogin luonne on siis kyettävä välittämään myös uudelle lukijalle. Referointitehtävä voidaan tehdä hyvinkin vaikeaksi tai helpoksi, aina sen mukaan kuinka mutkikkaiden merkityssuhteiden kognitiivista ja tekstuaalista hallintaa lukiolaiselta katsotaan voitavan vaatia.

En halua tässä propagoida mitään tehtävämallia lukion päättökokeeseen. Varsinainen sanomani on se, että tehtävänlaatijoiden ja sensoreiden pitäisi voida selvittää paremmin itselleen ja opettajille päättökokeen tehtävien tavoitteet ja arvioinnin kriteerit. Tehtävien tulisi myös käytännössä jotenkin vastata sitä monipuolisuutta, joka lukion oppimääräsuunnitelmassa on ilmaistu. Lukion ohjelmaan kuuluvat niin elämyksellinen, selostava, pohtiva kuin vaikuttavakin kielenkäyttö. Ongelmana on se, että niitä ei ole kyllin selvästi määritetty.

Asetin edellä kyseenalaiseksi sen, että aine olisi käytännössä niin vapaa tehtävätyyppi kuin sen sanotaan olevan. Entäpä jos se olisikin? Pitäisikö aineenkirjoittajille taata enemmän vapauksia, mutta säilyttää koe muuten nykyisellään? Näin oppilailla olisi sekä koulussa että ratkaisevassa päättökokeessa aina mahdollisuus valita, miten ja mitä hän kirjoittaa.

Kaavoittamisen välttämisestä ja omintakeisuuteen rohkaisemisesta ei ole muuta kuin hyvää sanottavaa. Nykyisen oppimääräsuunnitelman henki on kumminkin sellainen, että oppilaan olisi harjoiteltava erilaisia kielenkäytön lajeja, tietäen mitä harjoittelee. Ei siis riitä, että lukiolaiselle annetaan aina mahdollisuus kirjoittaa mitä haluaa, vaan hänelle on myös tietoisesti opetettava joitakin kielenkäytön tapoja. Oppimääräsuunnitelma ei toteudu, jos joku voi koko lukioajan kirjoittaa vain elämystekstejä ja joku toinen vain referaatteja. Oppilaan on jouduttava erilaisiin kielen-

käyttötilanteisiin, jotka kumminkaan eivät saa olla liian spesifisiä - lukiohan ei ole suoraan mihinkään ammattiin valmistava oppilaitos. Mielipidekirjoitus ja referaatti ovat esimerkkejä tehtävistä, joita voi tietoisesti harjoitella ja joiden taitaminen ei kuulu vain joidenkin ammattien vaatimuksiin. Näillä kirjoittamisen lajeilla on myös kieliyhteisössä vahva traditio, eikä niiden harjoittelu siis tähtää pelkästään kokeeseen. Se että opetuksen tavoitteet edes joiltakin osin selkeytetään päättökokeessa, auttaa selkeyttämään niitä myös opetuksessa.

Kirjallisuutta

Kaipainen, R. & K. Keravuori & P. Nallikari & A. Särkkä (toim.), 1984. Ylioppilasaineita 1984. SKS.

Kauppinen, Anneli & Leena Laurinen, 1984. Tekstejä teksteistä. Muisti- ja tekstilingvistiikan sovelluksia asiatekstin referoinnin problematiikkaan. SKS.

KOMMENTTI ANNELI KAUPPISEN PUHEENVUORON ÄIDINKIELEN TAIDON
MITTAAMISESTA LUKION PÄÄTTÖVAIHEESSA

Heljä Kauhanen
Joensuun normaalikoulu

This paper deals with the development of the Finnish Matriculation Examination in Finnish written presentation. The corresponding examination types in Sweden, the Federal Republic of Germany, the Netherlands and France will be discussed and later on, compared with that of Finland.

Nykyisenlainen suomenkielinen ylioppilasaine kirjoitettiin ensimmäisen kerran v. 1874. Aiheita oli yhdeksän, joista kuusi historiasta, yksi antiikin kirjallisuudesta ja yksi Kalevalasta. Ylioppilastutkintoon ei kuulunut reaalikoetta. Siksi valtaosa aineen aiheista oli reaaliaineista 1920-luvulle asti, jolloin reaalikoe erotettiin omaksi kokeekseen. Kirjallisuusaiheet olivat pitkän aikaa Kalevalasta, koska se oli ainoa kouluissa käytettävissä oleva suomenkielinen kaunokirjallinen teos; sitä tutkittiin jopa kahdeksan lukukautta.

Ylioppilasaineen tehtävänasettelussa on 110 vuoden kuluessa tapahtunut vähäisiä muutoksia. Aiheiden määrä on noussut ensin kymmeneen, sitten kahteentoista ja lopulta neljääntoista. V:sta 1970 on kirjoitettu kaksi ainetta viikon väliajoin. Äidinkielenopettajain liiton ehdotuksesta ylioppilasaineen 14 aihepiiriä on määriteltä ja määrittelyä on jo kerran tarkistettukin. Aiheisiin on tullut mukaan avoimia otsikoita, ja kääntöpuolelle on v:sta 1974 alkaen sijoitettu 1-2 aiheeseen virikemateriaalia.

Ainekirjoitus oli tärkein oppikoulujen kirjoittamisen harjoittelumuoto. Jo 1880-luvulta on peräisin vaatimus, että vuodessa oli kirjoitettava yhdeksän ainetta, niin keskikoulu- kuin lukio luokilla. Kun luokat olivat suuria, äidinkielenopettajilla riitti kotityötä.

Entisajan opetussuunnitelmat olivat tavoitteiden ilmaisemisessa ja metodisilta ohjeiltaan niukkasanaisia. Sitä paitsi niitä ei

ollut yleensä opettajien nähtävissä. Kirjoittamisen opetus oli eräänlaista suullista perimätietoa. Kaarlo Niemisen Ainekirjoituksen opas vuodelta 1925 ohjasi vuosikymmenet opetusta. Siinä on yleisen kirjoittamisstrategian mukaisia ohjeita otsikon valinnasta, aineiden keruusta, jäsentelystä, aloittamisesta ja lopettamisesta. 1960-luvulla ilmestyi aineenaitekokoelemlia ja Vuokko Raekallio-Tepon ja Terttu Tupalan kirjoittamisen oppaat. Pääpaino oli yleisen kirjoittamisstrategian opettamisella, mutta tekstilajejakin mainitaan.

V:n 1982 uusi lukion opetussuunnitelma on muuttanut äidinkielenkin oppimääräsuunnitelmaa. Äidinkielessä on kahdeksan tarkoin sisältöltään määriteltyä kurssia. Lukioon on ilmestynyt neljä kilpailevaa äidinkielen oppikirjasarjaa. Kirjallisten harjoitusten lukumäärää ei ole määritetty. Sisällöissä ei puhuta enää ainekirjoituksesta, vaan eri kursseissa perehdytään erilaisiin tekstilajeihin. Opettajan työ painottuu kurssien etukäteissuunnitteluun ja uusien sisältöjen opettamisen opetteluun.

Syksyllä 1982 minulla oli jatko-opiskelulukukauteni aikana tilaisuus tutustua neljässä Euroopan maassa lukiota vastaavan koulun kirjoituksenopetukseen ja päättötutkintoihin. Tässä vertailun vuoksi hiukan tietoja niistä:

Ruotsista on ylioppilaskirjoitukset poistettu, mutta kouluylihallitus laati keskuskokeita (centrala prov), jotka ovat yhteisiä kaikille päättöluokkien oppilaille. Äidinkielen centralt prov koostuu tiettyä aihepiiriä käsittelevistä tehtävistä ja 10-12 sivusta taustamateriaalia (aihepiirejä esim. Teatteri ja teatterikeskustelu, Eläminen ennen ja nyt, Väkivaltaviihde, Me ja kirjallisuutemme). Kaikissa tehtävissä on annettua materiaalia käytettävä hyväksi. Tekstilajeiksi ehdotetaan selostavaa tai pohdinta-aihetta, keskustelun johdantoa tai lisää, debattiartikkelia, pääkirjoitusta tai yleisönosastokirjoitusta, kirjettä tai novellia. Jokaiseen otsikkoon liittyy selostava johdanto. Äidinkielen opettaja korjaa kokeet alustavasti yksin, lopullisesti tehtävään määrätyn kollegansa kanssa. Tiettyinä päivinä syntyneiden oppilaiden kirjoitukset lähetetään kouluylihallitukseen. N. 1 kk:n päästä opettaja saa kouluylihallituksen toimittaman n. 50-sivuisen ohjevihkon kokeiden lopullista korjaamis-

ta helpottamaan. Kouluylihallitukseen koko maasta kertyneen otoksen avulla on määritelty, mitkä seikat ovat olennaisia eri otsikoista kirjoitettaessa. Vihko sisältää näytteitä eritasoisista suorituksista. Arvosanojen hajonnasta on vihkossa malli.

Ruotsalaisen koemallin etuna on, että oppilaalla on valmista materiaalia käytettävänä kirjoitustilanteessa, että hän saa vielä lisävihjeitä kunkin otsikon käsittelyä varten ja että hänen on myös mietittävä, minkä tekstilajin hän valitsee aineensa muodoksi. Ikuista väittelyä herättää, onko kunkin vuoden aihepiiri kaikille tuttu ja kirjoittamaan innostava. Oppilaalta menee melko suuri aika koeajasta tekstien lukemiseen. Korjausohjeet auttavat varsinkin aloittelevaa opettajaa.

Saksan Liittotasavallassa päättötutkinto on erilainen kussakin osavaltiossa. Tutustuin vain Hampurin päättötutkintoon. Hampurissa opettajat laativat itse oppilaittensa päättökoetehtävät. Esim. äidinkielen kirjallista koetta varten opettaja laatii luokalleen 5-6 tehtävää ja lähettää ne kouluvirastolle, joka valitsee niistä kolme vaihtoehtoa kokeeseen. Tehtävät ovat kielen ja kirjallisuuden alueelta ja liittyvät koulussa käsiteltyihin asioihin. Opettaja korjaa suoritukset itse ja joku kollega käy ne myös läpi. Baijerissa on keskitetty ylioppilastutkinto: kirjallisten kokeiden tehtävät ovat koko valtiossa samat.

Hollannissa päättötutkinto koostuu viimeisen kouluvuoden aikana suoritettavista koulun loppukokeista ja valtakunnallisista kirjallisista kokeista; lopulliseen keskiarvoon lasketaan puolet kummastakin. Meidän ylioppilaskirjoituksiamme vastaa suunnilleen VMÖ:n päättötutkinto, jonka kuitenkin suorittaa vain n. 10-15 % ikäluokasta. Valtakunnallisia kirjallisia päättökoetyyppejä äidinkielessä ovat vaikean abstraktin tekstin referointi ja ainekirjoitus. Referaatin korjaavat oma opettaja ja vieraalta paikkakunnalta tai vierasta koulusta tuleva toinen opettaja yhdessä, ainekirjoituksen oma opettaja yksinään. Referaatin korjaamisen avuksi opettajille tulee valtakunnallinen ohje, jossa on yleisiä ohjeita sekä kyseisestä tekstistä rakennekaavio, kustakin jaksosta mukaanotettavat tärkeimmät asiat ja niiden väliset suhteet. Valtakunnallisen kokeen lisäksi on koulun omia suullisia ja kirjallisia kokeita erityisesti kirjallisuudesta.

Ranskassa on valtakunnallisessa äidinkielen kirjallisessa ko-
keessa kolme vaihtoehtoa: Ensimmäinen vaihtoehto on annettu n. 700
sanan pituisen tekstin tiivistäminen neljäsosaansa; lisäksi on teks-
tistä annettava sanaselityksiä ja tekstistä poimittava kysymys,
josta on esitettävä kriittinen pohdinta. Toinen vaihtoehto on com-
mentaire composé eli erittäin tarkka tekstianalyysi annetusta kir-
jallisesti arvokkaasta tekstistä. Tämänäntyyppistä tarkkaa kirjalli-
suusanalyysia on ranskan tunneilla paljon harjoitettu. Kolmannessa
vaihtoehdossa annetaan kokelaalle jonkun tunnetun henkilön filosofin
mietelmä tai väite, jota kokelaan on käsiteltävä lukemansa kir-
jallisuuden pohjalta.

Ranskalaiset lukion äidinkielen opettajat kertoivat, että useim-
mat oppilaat valitsevat ensimmäisen vaihtoehdon, kirjallisesti har-
rastuneet toisen ja aniharvat kolmannen, koska se vaatii suurta luke-
neisuutta ja on erittäin vaikea. - Valtakunnallisten kokeiden lisäk-
si on koulun omia suullisia ja kirjallisia kokeita, joihin sisältyy
runsaasti kirjallisuutta.

Jokaisen maan koulujärjestelmä on historiallisen kehityksen tu-
los, eikä päättökokeita voi verrata toisiinsa tuntematta opetussuun-
nitelmia ja kouluun tulevien oppilaiden tasoa. Yhtä piirrettä ei voi
irrallaan siirtää maasta toiseen. Kuitenkin koetyyppejä vertailtaes-
sa erottuu muutamia nykyiselle suomalaiselle ylioppilasaineelle tyy-
pillisiä piirteitä:

- Koetyyppi on sama kumpanakin koepäivänä.
- Aiheet ovat monien oppiaineiden aloilta; muualla aiheet ovat
kielen ja kirjallisuuden tai yleistiedon alueilta.
- Aiheena on usein vain yksi rivi. Lisäohjeita ja virikemateriaa-
lia on vähän.
- Korjaus on kaksinkertainen: aineet korjaa ensin opettaja ja
sitten ylioppilastutkintolautakunnan sensori. Opettaja ei saa
korjausohjeita, vaan ohjanta ja korjausohjeet suunnataan yli-
oppilastutkintolautakunnan sensoreille.
- Koulun kokeet ja ylioppilastutkinto on erotettu toisistaan.

Suomessa toivoisin äidinkielen päättötutkinnon kehittyvän sekä
perinteisen yleisen kirjoittamisen eli ainekirjoituksen kehittämisen

että tekstilajien harjoittelun suuntaan. Tekstilajeille on osoitettavissa yhteisiä piirteitä ja niitä voi oppia ja opettaa kirjoittamaan. Tekstilingvistiikka voi hyödyntää äidinkielen opetusta. Tarvitaan kuitenkin kokeilua ja tutkimusta. Anneli Kauppinen jo laatimat raportit ovat erittäin hyödyllistä opettajien koulutusmateriaalia; mielihiteen kirjoittamista koskeva raportti toivottavasti valmistuu vielä. Oppimateriaaleja on kehiteltävä ja opettajien jatkokoulutusta lisättävä. Erittäin suotavana pidän kirjallisuuden arvostuksen lisäämistä ja kirjallisuusanalyysin mahdollista ottamista päättötutkintoon. Pienimuotoista kirjallisuusanalyysia ei Suomessa kuitenkaan kirjallisuudenopetuksessa ole tähän asti kehitelty. Kaiken kaikkiaan päättötutkinnon kehittäminen antaisi lukion äidinkielenopetukselle selvän tavoitteen.

KOULULUOKAN DISKURSSIN PIIRTEITÄ

Kyllikki Keravuori (Tampereen yliopisto)

The article explores classroom discourse at the upper level of the Finnish comprehensive school. The observations are based on some biology, religion and mother tongue classes of a seventh class. The lessons turn out to be semantically coherent and structurally logically planned. The interaction during the lessons is, due to the chosen teaching method, teacher-centered. The interaction of the teachers mainly occurs so that they ask clearly restricted questions (aiming at a certain answer). Thus, the answers given by the students are brief. Because of the brevity of the answers it is not possible to analyze the students' personal characteristics in using the language. The teacher-centered lesson only describes one part of classroom discourse. As to the students the picture will be different when other teaching methods are employed.

1 Taustaa

Äidinkielenopettajana minua on vuosien varrella alkanut

erityisesti kiinnostaa päivittäisten koulukeskustelujen kieli. Kasvatustieteellistä vuorovaikutuksen tutkimusta on tehty parinkymmenen vuoden ajan. Meillä lienevät yleisimmin tuttuja Flandersin ja Bellackin tutkimukset, ja varsinkin Bellackin ja hänen työtovereittensa ajatuksiin pohjaten on Helsingissä toteutettu laaja didaktisen prosessianalyysin tutkimushanke, DPA-Helsinki. Mainitut tutkimukset eivät ole olleet kielen tutkimuksia. On pyritty selvittämään luokan ja opettajan välisiä valtasuhteita, osapuolten puheen määriä, opettajan vaikuttamistapoja ja vuorovaikutusprosessia oppitunnilla. Välineenä analyyseissa on käytetty mm. kieltä.

Jos on kiinnostunut siitä, millaista on oppitunnin kieli, joudutaan kielen funktioitten havainnointiin. Tarkoitan täällä samaa asiaa kuin Hakulinen ja Karlsson, kun he Nykysuomen lauseopissa erottavat lausetyypit ja puhefunktiot toisistaan. Yhdellä lausetyypillä saatetaan ilmaista erilaisia puhefunktioita. Kukin puhefunktio toteutuu ensisijaisesti jollakin lausetyypillä, kuten toteamus väitelauseella, käsky imperatiivilauseella ja kysymys kysymyslauseella. Mainitut lausetyypit esiintyvät myös toissijaisesti muissakin puhefunktioissa.

Esimerkkejä toissijaisesta käytöstä:

1. nyt kaikki keskittyy, liikkumattomuusleikki
(väitelause, jota on käytetty käskynä)
2. niin te vaan kirjoitate siihen paperille sen oman
ajatuksen (väitelause käskynä)
3. kuinka kauan mun pitää oottaa ennen kun täällä
rauhotutaan (kysymyslause, jota on käytetty käskynä)
4. eks sä ymmärtäny mitä mä sanoin (kysymys käskynä)
5. nyt te ravustajat tiedätte ainakin sen että mihin
vuorokauden aikaan rapu on liikkocellä (väitelause,
jota on käytetty kysymyksenä)
6. täs on yks asia mikä saattaa tuntua merkilliseltä
(väite kysymyksenä)

7. tehkää ehdotuksia 'miten kirjoitatte' (käsky-
lause, jota on käytetty kysymyksenä)
8. koettakaa saada nuo toistuvuudet pois 'millä sa-
nalla korvaisitte' (käsky-lause kysymyksenä)
(Kaikki esimerkit ovat nauhoittamistani oppitun-
neista.)

Tärkeää oppitunnilla on, miten puhujat, opettaja ja oppilaat, tunnistavat, mitä sanottu kulloinkin todella tarkoittaa.

2. Aineisto ja analyysimenetelmä

Aineisto, josta olen tuntidiskursseja seurannut ja eritellyt, on koottu peruskoulun yläasteelta, 7. luokalta. Edusteilla on kolme oppiainetta: biologia, uskonto ja äidinkieli. Ajatukseni oli, että näiden aineiden tunneilla saataisiin mahdollisimman paljon talteen oppilaittenkin puhetta. Tunnit olivat sopimuksen mukaan opettajajohtoisia. Aineisto on kuvanauhoitettua.

Oppitunteja eritellessäni olen käyttänyt hyväkseni Sinclairin ja Coulthardin funktionaalista menetelmää, diskurssianalyysia. Se on kehitetty puheaktiteorian pohjalta.

Jotta oppitunnin keskustelua voidaan ymmärtää, pitää ottaa huomioon puhujien, siis opettajan ja oppilaitten, yhteiset tiedot ja sosiaaliset sopimukset eli mahdollinen olemassa oleva ei-kielellinen tieto. Sitä on yleensä tieto koulusta, tieto luokan tavoista, tieto nimenomaisesta tarkkailtavasta tunnista tai hetkestä. Tätä pragmaattista tietoa tarvitaan, jotta kävisi ymmärrettäväksi, miten yksityinen ilmaus toimii diskurssiyhteydessään. Tekstintutkija Widdowson sanookin, että normaali kielellinen käyttäytyminen ei ole erillisten lauseitten tuottamista, vaan se on diskurssin rakentamista ilmauksista.

Ymmärtämisen kannalta on tärkeää tietää, millaiset diskurssiyksiköt seuraavat toisiaan ja ovat suhteessa toinen

toiseensa. Toisin sanoen ilmauksen paikka diskurssissa on merkityksellinen, ei pelkkä muoto.

Diskurssianalyysi on hierarkkinen menetelmä. Siinä on viisi porrasta - ylimmästä alimpaan: tunti, episodi, vuoropari, siirto ja akti. Pienimmällä yksiköllä ei ole määriteltävää rakennetta, mutta muiden yksiköiden rakenne voidaan määritellä lähinnä alapuolellaan olevalla yksiköllä. Seuraava kuvio havainnollistaa portaittaista rakennetta.

tunti



episodi



vuoropari



siirto



akti



3 Diskurssin semanttiset yksiköt

Tunti on koulutyössä ajan rajaama yksikkö. Riippumatta oppiaineesta tai opettajasta jokaisella tunnilla on teema, joka johdetaan opetussuunnitelmasta. Teeman käsittely jakaantuu - oppimisen tähden - alateemoiksi. Nämä alateemat ovat semanttisessa suhteessa koko tunnin teemaan. Tunti on siis merkitysyksikkö.

Voisi olettaa, että opettajalla ja oppilailla on mielessään jonkinlainen tuntisäännöstö, jolla he tulkitsevat tunnin jäsentymistä. Tällöin he käyttävät myös ei-kielellisiä kykyjään, kuten muistia ja aivojen tiedonprosessointia. Kielellisellä puolella on havaittavissa hierarkkinen jäsentyminen, jota hallitsee teema.

Tunnin rakennetta voidaan kuvata tunnin alapuolisella yksiköllä, siis episodilla. (Ks. edellä esitettyä rakennekuviota.) Voidaan esimerkiksi kysyä, monestako episodista tunti koostuu. Minun aineistossani episodimäärä vaihteli 9 - 24.

Tutkimani tunnit olivat semanttisesti hyvin koherentteja. Erityisesti näin oli asianlaite biologian tunneilla ja lähes samoin äidinkielen tunneilla. Uskonnon tuntien kulkua teema ei sitonut yhtä tiukasti. Epäkoherenssia voivat olla työrauhapoikkeamat, erilliset järjestelyt ja sopimukset ja jotkut hyvin löyhästi teemaa sivuavat episodit.

Jos tavoitteena pidetään puhdasta tiedonsiirtoa, koherentti tunti on hyvä. Voidaan kuitenkin kysyä, mitä semanttiseen koherenssiin johtava käsittelytapa merkitsee oppilaan itsenäisen ajattelun kehittymiselle. Opettaja johtaa keskustelua - kyselyä niin tiukasti, ettei harhavastauksia synny; usein oppilaan ei tarvitse ajatellakaan, kunhan vain tietää.

Jokaisella tuntiin sisältyvällä episodilla on oma teemansa, joka on tavallisesti sidoksissa koko tunnin teemaan. Siis myös episodi on merkityksen yksikkö. Rakenteellisiakin tuntomerkkejä episodilla on. Episodin alusaa on opettajan puheenvuoro eli rajavuoropari, joka esittelee episodin teeman. Se on hyvin säännönmukaista, sillä aineistoni 103 episodista 94 alkoi rajavuoroparilla. Vastaavasti episodin lopussa on rajaus. Opettaja sanoo, mihin teeman käsittelyssä on päädytty. Lopun rajaus ei ole yhtä säännönmukaista kuin alun, sillä rajavuoropari esiintyi 54 tapauksessa 103 mahdollisesta.

Episodin sisältö rakennetaan opetusvuoroparein. Opettaja ja oppilaat käsittelevät rajauksessa ilmoitettua teemaa. Aineittain oli eroja käsittelyn kestossa. Uskonossa oli opetusvuoropareja episodissa keskimäärin 6, biologiassa yli 10 ja äidinkieliessä peräti 12. Yhtä alateemaa siis käsiteltiin aineittain eri pitkiä aikoja. Todennäköisesti tunnin teema säätelee episodien pituutta. Jos uskonnotunnilla on aiheena "maailman luominen", tuntuu luonnolliselta, että episodit eivät monesta opetusvuoroparista rakennu. Seitsemäsluokkalainen tuskin voi kovin kauan puhua, ei riitä kuvittelukyky eikä tieto; ei sellainen tilanne ole myöskään ominaan lisäämään keskusteluhaluutta.

Biologian tuntien aiheet taas ovat monesti tuttuja oppilaiden omasta havaintopiiristä - kuten nauhoitettujen tuntien aihe "hauki" - joten keskustelu on käytännössä mahdollista. Aineistoni äidinkielen tuntien teemoina olivat "yhdyssanat" ja "lauseitten sitominen tekstiksi". Näiden aiheiden käsittely ei tuonut seitsemäsluokkalaistille varsinaisesti uutta tietoa, vaan sen sijaan pyrittiin vahvistamaan ennen opittuja asioita. Näin ollen kussakin alateemassa saatettiin pysytellä monen vuoroparin ajan.

On myös mahdollista, että opettajan persoonallakin on vaikutusta episodin pituuteen, mutta minun tutkimukseni valossa sellaista ei voi väittää, saati todistaa.

4 Diskurssin interaktiivinen taso

4.1 Vuoropari

Edeltävässä jaksossa olen selittänyt episodien kestoa vuoropari-nimisellä diskurssiyksiköllä. Vuoropari on kouludiskurssin interaktiivinen yksikkö. Vuoroparin tasolla voidaan tarkkailla, miten vuorovaikutus sujuu opettajan ja oppilaan välillä koulutunnilla.

Rajavuoropari on pelkästään opettajan puheenvuoro, jota hän käyttää diskurssin jäsentämiseen. Eri opettajat näyttävät jäsentävän

eri suuressa määrin. Minun aineistossani uskonnon opettaja jäseni eniten. Esitän arvelunani, että asiaan vaikuttavat oppiaineen vaikeus, teema sekä opettajan persoonakin.

Varsinainen opetusvuoropari koostuu kolmesta osasta: opettajan aloituksesta, oppilaan vastauksesta ja opettajan seuruupuheenvuorosta. Vuoropari on siis periaatteessa kolmiosainen, vaikkakaan seurua ei säännöllisesti esiinny.

Funktionensa mukaan vuoroparit ovat kolmenlaisia: kyselyjä, ohjailuja tai informointeja. Koko aineistossani funktion mukainen jatkautuminen oli seuraava:

kyselyvuoropareja	466
ohjailuvuoropareja	180
informointivuoropareja	185

Tulos kertoo, että kaiken kaikkiaan näillä tunneilla kyselevä vuoro vaikutuksen tapa oli ehdottomasti suosituin.

Kyselyjä voi myös luokitella niiden tavoitteen mukaan: 1) tavallinen kysely, jolla perätään tietoa, havaintoa tms., 2) tarkistus, jossa otetaan selvää, onko asia ymmärretty, 3) uudelleenaloitus, johon on ryhdyttävä, ellei tavalliseen kyselyyn ole saatu minkäänlaista vastausta, 4) vastauksen kehittäminen, jossa yritetään vajavainen vastaus kääntää diskurssin hyödyksi, 5) vastausten kerääminen, jossa taktisesti yritetään kerätä useita vastauksia, ennen kuin niitä mitenkään kommentoidaan, 6) toiston pyytäminen, joka esitetään tavallisimmin silloin, kun puheenvuoro on ollut kuulumaton. Toiston pyyntö on mahdollisesti myös taktisista syistä käytetty kyselyn muoto.

4.2 Kysymisen strategiat

On ilmeistä, että opettajilla on erilaisia kysymisen strategioita. Uskonnotunnilla syntyi joka kolmannen kyselyn jälkeen tilanne, että vastausta ei tullut. Silloin opettaja ajautui uudelleenaloituksiin. Biologian opettaja sen sijaan oli samassa tilanteessa joka 9. kyselyn

jälkeen. Hän joutui aloittamaan uudelleen yleensä siitä syystä, että hänen tekemänsä kysymys olisi edellyttänyt tarkkaa asiantietoa, mutta oppilaat eivät sitä tienneet tai eivät tohtineet sanoa mitään. Uskonnon tunnilla vastaukseton kysymys olisi edellyttänyt omia havaintoja tai päättelyä, toisaalta myös tosioita niin kuin biologiassakin.

Äidinkielen opettaja johti oppilaita kysymyksillään päätelmiin. Hän teki sitä hyvin pienin askelin, niin ettei oppilaiden keskittymiskyky aina riittänyt loppuun asti. Uudelleenaloitukseen äidinkielen opettaja joutui kuitenkin ani harvoin eli vasta joka 10. kyselyn jälkeen.

Opettajat käyttivät saamansa vastauksen eri tavoin hyväkseen. Jos vastaus oli väärä tai toisenlainen, kuin opettaja mielessään oli kuvitellut, riippui opettajan strategiasta, miten diskurssi kehittyi. Innokkain vastausten hyväksikäyttäjäksi oli äidinkielen opettaja. Noin joka 4. kyselyn hän aloitti kehitelläkseen oppilaan vastauksen diskurssin hyväksi. Menettelyn ero oli suuri verrattuna biologian ja uskonnon opettajaan. Biologian opettaja kehitti vasta joka 8. vastausta ja uskonnon opettaja ryhtyi siihen äärimmäisen harvoin.

Äidinkielen ja uskonnon opettaja tarkistivat tunnin sujumista ja oppilaiden ymmärtämistä. Sen sijaan biologian opettajan strategiassa se oli harvinaista.

4.3 Vastaukset ja niiden kommentointi

Vuoroparin oppilaan osuutta ovat vastaukset. Oppilaat vastasivat huomiota herättävän usein. Kaikista aloituksista 85 %:iin tuli vastaus. Samoin seuruu eli opettajan kielellinen kommentti vastaukseen tuli 60 %:iin tehdyistä kyselevistä aloituksista. Seuruiden runsaus oli iloinen yllätys. Meikäläisessä didaktisessa kirjallisuudessa näet annetaan toisenlaisia ohjeita ja aivan ilmeisesti vastoin kielen käytäntöä ja luontevia keskustelutapoja.

Aineiden välisiä eroja näkyi vastausten saannissa ja kommentoinnissa. Biologian ja äidinkielen opettaja saivat vastauksia kysymyksiinsä lähes 90-prosenttisesti, mutta uskonnon opettaja 60-prosenttisesti. Samalla on tosin sanottava, että uskonnon

opettaja ei erityisesti suosinut kyselevää interaktion tapaa, vaan hän melkein yhtä tasaisesti myös ohjasi ja informoi. Sen sijaan äidinkielen ja biologian opettaja harjoittivat voittopuolisesti kyselevää vuorovaikutusta. Myös vastausten kommentointia uskonnon opettaja suoritti vähemmän kuin toiset kaksi.

5 Yksilöllisiä piirteitä

Esiteltä diskurssin vuorovaikutustaso koostuu rakenteellisesti siirroista. Yksityinen siirto merkitsee oppitunnilla puheenvuoroa. Opettajajohtoisella tunnilla avaus- ja seurussiirrot ovat opettajan ja vastaussiirrot oppilaan puheenvuoroja. Avaukset ovat luonteeltaan joko kysyviä, kehottavia tai informoivia. Se tarkoittaa, että puheenvuoroissa on joko kysyvä, kehottava tai informoiva ydinosa. Sen lisäksi niissä on tavallisesti myös muuta ainesta. (Ks. diskurssin rakennekuviosta porrasta siirto.)

Esimerkkejä: no

mikä evä tuossa on

Jukka

sieltä tuli uteliaisuus, rohkeus ja viisaus,
niin mihinkä ihminen nyt ihan selvästi pyr-
kii kaikella tieteellään ja tutkimisellaan
ja toiminnallaan ja rohkeudellaan, mihin
ihminen pyrkii

Tutkimieni tuntien avauspuheenvuoroissa oli keskimäärin 2,2 aktia eli ytimen lisäksi aina jotakin muutakin.

Puheenvuoroissa näkyvät opettajien persoonalliset diskurssipiirteet. Eroja syntyy mm. siitä, kuinka monesta osasta ja minkälaisista osista opettajat rakentavat puheenvuoronsa. Puheenvuoron pituus - aktimääräinen - on siis yksi erotteleva tekijä. Toiseksi myös se erottelee, suosiiko opettaja kyselemistä, ohjailemista vai informoimista vai ovatko kaikki tavat keskustelussa tasapuolisesti edusteilla. Eli interaktiotapa kieltii persoonallisista valinnoista.

Oppilaan vastausiirrot eivät aineistoni mukaan kuvastele lainkaan persoonallisia piirteitä. Oppilaat vastasivat opettajan kysymyksiin yksiaiktisella vastauksella eikä aineittain ollut mitään eroa.

Esimerkkejä:

kysymys: minkä kokosia ne teiän hauet on ollu

vastaus: parikilosia (biol.)

kysymys: te ootte nähny täällä lokkeja, koiran ja sikoja

miten ihminen eroo eläimestä muulla tavalla kun siinä että että ihminen panee sen remmin koiran kaulaan eikä eikä koira ihmisen kaulaan, miten muuten ihminen eroo

vastaus: ulkomuodosta (usk.)

Edellä kuvattu vastaanmistapa johtaa tarkastelemaan kysymysten ja vastausten keskinäistä riippuvuutta. Keskustelussa vastaaja tavallisesti sanoo kysytyn asian eikä toista kaikkea mitä kysymyksessä on ollut esillä. Muodon mukaan kysymyslauseet jaetaan kahtia: vaihtoehto- ja hakukysymyksiin. Esimerkki vaihtoehtokysymyksestä:

kummalla näistä vaihtoehtoista, tiikerillä vai kääpiömyskihirvellä, on enemmän onnenpipanoita? (äk.)

Vaihtoehtokysymyksiin kuuluvat myös ns. -ko-kysymykset, joihin voidaan vastata myöntävästi tai kieltävästi.

korvaako se vai ei? (äk.)

Hakukysymyksellä opettaja taas pyytää tietoa kysymyksensä yhteen osaan. Etsitty tieto voi olla selvästi rajallinen, mutta se saattaa toisaalta olla alueeltaan epämääräinenkin. Juuri hakukysymyksiin voidaan odottaa lauseen mittaisia vastauksia. Vastauksen laajuus riippuu siitä, miten täsmällistä tai rajallista tietoa kysyjä hakee.

Esimerkkejä:

kysymys: mikä se uloke on nimeltänsä, joka
sieltä voi tulla esiin

vastaus: jalka

kysymys: millä rapu hengittää

vastaus: kiduksilla

kysymys: mitä sille kuorelle tapahtuu

vastaus: se kasvaa

kysymys: miten sudenkorento eroaa nössänään
kärpäsestä paitsi että se on sitä
suurempi

vastaus: sil on isommat siivet

Kysymysten yhteisenä piirteenä oli, että ne olivat luokiteltavissa hakukysymyksiksi. kysymykset olivat alaltaan hyvin spesifejä, lukuun ottamatta uskonnon tunnin kysymyksiä. Spesifit kysymykset pakottavat oppilaat vastaamaan lyhyesti. Oppilas täydentää opettajan kysymyksen puuttuvan tiedon, mutta kysymyksen perusteella hänellä on harvoin mahdollisuus ilmaista omia ajatuksiaan. Minun aineistossani uskonnon opettaja oli tässä suhteessa erilainen. Hän teki eniten väljiä kysymyksiä.

Oppilaiden vastaukset olivat yleensä joko yksisanaisia tai lausekkeen mittaisia. Lauseen pituisia vastauksia oli kaikista vastauksista noin 10 % verran.

Esimerkkejä:

miten selitätte sen että vesimittari pysyy
siinä pinnalla?

sil on ilmatyynyt jaloissa (biol.)

miten yhdistettäisiin nämä kaksi yhdeksi virkkeeksi?

laitetaan tuohon huomasi-sanan perään pilkku ja sitten kirjoitetaan eteenpäin (äk.)

minkälainen käsitys tuosta raamatun alusta nyt sitten teille muodostuu että miksi jumala tän loi?
että erottaa yön ja päivän (usk.)

Lauseen pituisia vastauksia näyttää tulevan hakukysymyksiin, jotka alkavat miten, millä tavalla, miksi -kysymyksenalla. Oppilaille jää tällaisiin kysymyksiin vastatessaan itsenäistä liikkumatilaa, joten vastauskin on pidempi.

Täsmällisiä hakukysymyksiä tekevä opettaja pyrkii tunneillaan oletettavasti käsittelemään mahdollisimman paljon ja tehokkaasti tietoja. Epätäsmälliseltä vaikuttava kysyminen antaa oppilaan omalle ajattelulle ja ilmaisulle enemmän tilaa, mutta tunti ei ehkä suju tiedon siirtämisen kannalta etukäteissuunnitelman mukaan ja tunnin teeman käsittelyyn saattaa tulla poikkeamia.

Oma-aloitteista oppilaan puhetta oli vähän. Oppilaan aloittamia vuoropareja oli kaiken kaikkiaan reilusti alle 10 %:n. Oppilaat kysyivät menettelytapoja, ja niihin opettajat olivat valmiit vastaamaan.

Esimerkkejä:

millä sivulla se on?
tuodaanks saappaat?
saaks tähän paperiin kirjottaa?
kuinka paljon siihen täytyy jättää tilaa?

Tunnin teemaan liittyviä asioita oppilaat kysyivät harvoin, mutta kylläkin joskus. Monesti tällainen kysymys jää opettajalta myös kuulematta, ja toisinaan syntyy vaikutelma, että opettaja on kuuntelussaan selektiivinen.

Esimerkkejä teemaa koskevista kysymyksistä:

onks toi suu?

ope, mitä tapahtuu jos laittaa kiven elävän
simpukan sisään?

mihkä pilkku pannaan jos jos tulee rivin al-
kuun?

6 Miten eteenpäin?

Esittämäni havainnot pätevät opettajajohtoiseen koululuokan diskurssiin. Se on kuitenkin vain yksi puoli kunkin luokan todellisesta diskurssista. Askarruttaa, millaista on nimenomaan oppilaan kieli toisenlaisia työmenetelmiä käytettäessä, jolloin opettaja ei pääse samalla tavalla vaikuttamaan oppilaitten puheen kulkuun.

Peruskoulun tulosta lähtien on yritetty monipuolistaa koulun työtapoja. Opettajaajohtoisten tuntien yksinvaltiutta on tavallisimmin horjutettu ryhmätyöllä. Oppilaiden sosiaalistamiseen se työmuoto soveltuu hyvin. Nyös tiedonhankintaan ryhmätyötä voi käyttää, mutta koulun on suostuttava hyväksymään se, että tarvitaan aikaa, jotta tiedonhankinta onnistuu.

Joka tapauksessa ryhmätyötä käytetään nykyään melko runsaasti. Olen jonkin verran nauhoittanut samojen aineiden ryhmäkeskusteluja ja ryhmätyörupeamia. Arvattavissa jo oli, että kuva oppilaan puheesta käy toisenlaiseksi kuin opettajaajohtoisilla tunneilla.

Puheenvuorot näyttävät kautta linjan olevan pidempiä. Itse keskustelu vaikuttaa koherentilta siitä huolimatta, että kuvanauhoitus saattaa aiheuttaa joillekuille seitsemäsluokkalaistille voimakasta esiintymisen tarvetta.

Vaikutelmat ovat toistaiseksi pinnallisia, koska analyysia ei ole tehty. Mutta kuitenkin on varmaa, että tarvitaan tietoa myös ryhmäkeskustelutuntien kielestä, ennen kuin voidaan puhua oppilaan kielestä kouludiskurssissa.

Lähteet

Keravuori, Kyllikki, 1984. Opettajajohtoisen opetuskesustelun kielellisiä piirteitä. Tampereen Opettajankoulutuslaitoksen julkaisuja. Tampereen yliopisto.

Sinclair, J. McH. & R.M. Coulthard, 1975. Towards an Analysis of Discourse. The English used by teacher and pupils. London: Oxford University Press.

WISSEN UND KÖNNEN IM ERWERB DER DRITTEN UND VIERTEN FREMDSPRACHE

Matti Luukkainen (Universität Helsinki)

Abstract

Infolge der Reform der gymnasialen Oberstufe ist die Anzahl der Schüler, die Deutsch, Französisch oder Russisch als dritte oder vierte Fremdsprache wählen können, rapide zurückgegangen. Zugleich wurde das verfügbare Zeitpensum stark gekürzt. Deshalb ist bei der Planung des Unterrichts in diesen Sprachen besondere Sorgfalt geboten. Der Lernstoff muss stark reduziert werden. Was in den Unterricht mit einbezogen wird, muss so gut geübt werden, dass bei den Schülern ein Gefühl der Geborgenheit entsteht. Einfachheit und Flüssigkeit gelten als Ziel. Aufgabe der Schule muss es sein, eine gute Motivation zu schaffen und das Sprachgefühl positiv zu entwickeln.

In erfreulichem Ausmass stimmt der Geist der Beiträge in diesem Symposium miteinander überein - unabhängig davon, ob es sich um die Probleme der Fremdsprache oder der Muttersprache handelt. Neben den Ähnlichkeiten gibt es natürlich auch Unterschiede. Mein Beitrag bezieht sich vornehmlich auf die Probleme des Deutschen. Deshalb geht es im folgenden speziell um den Erwerb einer dritten bzw. vierten Fremdsprache und um den Unterricht der etwas gereiften Schüler der reformierten gymnasialen Oberstufe. Gerade auf diesen Tatbestand dürften einige Akzentuationen zurückzuführen sein, die von den anderen Ausführungen abweichen.

Es scheint unumgänglich zu sein, dass viele der Entwicklungen im Bereich menschlicher Tätigkeiten im Zickzack verlaufen. Das gilt auch für die Entwicklung der Sprachunterrichtsmethoden. Vor nicht allzu langer Zeit bestand der Sprachunterricht fast ausschliesslich aus Grammatik und Übersetzung. In der Folge waren der kontrastive Aspekt und jegliche intellektuelle Analyse aus den Sprachstunden verbannt. Es dauerte jedoch nicht lange, bis ein hemmungsloser Gebrauch der Muttersprache wieder erlaubt war. Bald werden die Phrasen und Strukturen in Form von Dialogen stur auswendig gepaukt, bald wird die Kommunikation als kognitiver Prozess verstanden. Für

die einen ist die Grammatik nach wie vor ein Schrecken, für die anderen ist ein angemessenes Angebot von Strukturen ein selbstverständlicher Aspekt auch im kommunikativen Fremdsprachenunterricht.

Jede Unterrichtsmethode birgt unverkennbare Vorzüge, aber auch Nachteile in sich. Deshalb ist die orthodoxe Anwendung irgendeiner einzelnen Methode grundsätzlich abzulehnen. Wenn man von der einen Methode zu der anderen übergeht, soll man also jeweils nicht das Kind mit dem Bade ausschütten. Ein idealer Sprachunterricht baut auf den wesentlichen Erkenntnissen der verschiedenen Methoden auf, wenn sie sich nicht widersprechen. Im Hinblick auf die Lernergebnisse ist jedoch die Art der Wechselwirkung zwischen der Person des Lehrers und der jeweiligen Schülergruppe entscheidend. Letzten Endes kommt es also darauf an, wie der einzelne Lehrer imstande ist, seine Schüler zum Spracherlernen zu motivieren. Dabei muss die angewandte Methode mit der Persönlichkeit des Lehrers im Einklang stehen. Extrem vorprogrammierte Lehrgänge, wie sie teilweise auch schon für unsere gymnasiale Oberstufe vorliegen, betrachte ich aus dem Grunde etwas skeptisch.

Wenn wir heute in Finnland über das Erlernen der dritten bzw. vierten Fremdsprache sprechen, meinen wir hauptsächlich das Deutsche, das Französische und das Russische. Und wenn es um die mit dem Unterricht in diesen Sprachen verbundene Problematik geht, können wir nicht umhin, auf das sprach- und kulturpolitische Unglück einzugehen, das daraus folgte, dass seit Mitte der 60er Jahre der Unterricht in Englisch in der 3. Klasse der Gesamtschule eingeführt wurde. Als Preis dafür, dass in der heutigen Gesamtschule sowohl Englisch als auch Schwedisch in luxuriöser Weise als zwei verschiedene Lehrgänge von unterschiedlicher Dauer (A-Sprache, B-Sprache) gelehrt werden, sind Deutsch, Französisch und Russisch in die Position von seltenen exotischen Sprachen geraten. Im Widerspruch zu den konstruktiven Vorschlägen des Sprachenprogrammkomitees von Numminen ist der Unterricht in Deutsch, Französisch und Russisch seit 1978 rapide zurückgegangen. Infolge der Reform der gymnasialen Oberstufe wird derselbe unglückliche Trend in Zukunft noch weitergehen.

Bei der Planung eines angemessenen Fremdsprachenangebots erwägt das Sprachenprogrammkomitee von 1978 insgesamt acht verschiedene Modelle. Unter den Vorschlägen fehlt nur eine Möglichkeit, die meiner Ansicht nach allerdings die naheliegendste Lösung wäre: Wenn wir in Finnland davon ausgehen, dass jeder Bürger die zweite Landessprache in der Gesamtschule lernen muss, sollte jedem Schüler als A-Sprache Schwedisch bzw. Finnisch angeboten werden. Wenn den etwas älteren Schülern Englisch, Deutsch, Französisch oder Russisch als B-Sprachen angeboten würden, wäre die angestrebte Diversifikation eher möglich als im heutigen System (vgl. Luukkainen & Schröder 1981).

Als die Gymnasialreform im Schuljahr 1982-1983 in der 1. Klasse der gymnasialen Oberstufe in Kraft trat, reduzierte sich die Zahl der Schüler, die eine C-Sprache oder D-Sprache erlernen, schon um ein Drittel. Für das Schuljahr 1983-1984, wo die Reform bereits für die beiden ersten Klassen galt, liegen folgende Statistiken vor (um Vergleiche zu ermöglichen, werden im folgenden auch die Zahlen für die 3. Klasse angegeben, wo die Reform noch nicht durchgeführt war; die Prozentzahlen beziehen sich auf die Gesamtmenge der Schüler auf der jeweiligen Stufe; die Angaben für Englisch, Spanisch, Italienisch, Latein und Lappisch werden wegen der geringen Zahlen nicht extra angeführt):

C-Sprache	1. Klasse		2. Klasse		3. Klasse	
Deutsch	11843	33,1%	11584	33,1%	12456	41,2%
Französisch	2174	6,1%	1985	5,7%	1579	5,2%
Russisch	769	2,2%	648	1,8%	874	2,9%
Insgesamt	15191	42,5%	14632	41,8%	15268	50,5%

D-Sprache	1. Klasse		2. Klasse		3. Klasse	
Deutsch	10929	30,6%	11531	32,9%	13038	43,1%
Französisch	4398	12,3%	3654	10,4%	5200	17,2%
Russisch	2078	5,8%	1924	5,5%	2363	7,8%
Insgesamt	18061	50,5%	17595	50,2%	21375	70,6%

Weil in der 3. Klasse der reformierten Oberstufe die Wahl einer C- bzw. D-Sprache auch für die Nichtmathematiker entweder nicht obligatorisch oder in vielen Fällen sogar überhaupt nicht möglich ist, ist sie laut inoffiziellen Statistiken im Schuljahr 1984-1985 von einer Hälfte der Schüler auch nicht gewählt worden. Die Anmeldungen für die fakultative Reifeprüfung in diesen Sprachen sind um ein Drittel zurückgegangen, wie aus den folgenden Angaben des Abiturausschusses hervorgeht (voraussichtlich werden ca. 85-90% der Angemeldeten an der Prüfung teilnehmen):

	Frühjahr 1984	Frühjahr 1985	Abnahme
Deutsch	17439	11826	32,2%
Französisch	3645	2834	22,2%
Russisch	1958	1277	34,8%

Zur Zeit werden für den Unterricht in den C- und D-Sprachen noch weitere Reduktionen erwogen, die besonders die Anzahl der für die letzte Klasse der gymnasialen Oberstufe vorgesehenen Kurse betreffen. Wir können also sehen, dass in den letzten Jahren hierzulande im Bereich des Unterrichts von Deutsch, Französisch und Russisch eine verhängnisvolle Umwälzung erfolgt ist. Gleichzeitig mit der Verringerung der Anzahl der Schüler, die diese Sprachen erlernen können, wird die für die restlichen Schüler verbliebene Lernzeit verkürzt. Es ist sehr zu bedauern, dass ausgerechnet die Sprachen, die aufgrund ihrer linguistischen Struktur die längste Übungszeit erfordern, im Sprachenprogramm der Schule die schwächste Stellung einnehmen. Diese Tatsache muss sich auf die Lernmotivation der Schüler negativ auswirken, was also schon jetzt in der abnehmenden Teilnahme am Abitur zum Ausdruck kommt.

Vor dem Hintergrund der oben umrissenen Entwicklung ist bei der Planung des Unterrichts in der dritten und vierten Fremdsprache besondere Sorgfalt geboten. Man muss dafür sorgen, dass die Arbeitsmethoden realistisch, ökonomisch und effektiv sind. Wenn ich

die Aussage meines Referats in einem Satz festhalten sollte, so würde ich sagen: Auf die Qualität kommt es an, nicht auf die Quantität. Mit anderen Worten: Der Lernstoff wird stark, aber sinnvoll reduziert. Was in den Unterricht mit einbezogen wird, wird so gut gelehrt, geübt und automatisiert, dass bei den Schülern ein Gefühl der Geborgenheit entsteht. Eine richtige Motivation und eine günstige Entwicklung des Sprachgefühls nehmen die Schlüsselposition ein (vgl. Harmia & Luukkainen 1977).

Wegen der Kürze der Lernzeit ist es nicht mehr möglich, im Rahmen des Unterrichts in den Dritt- bzw. Viertsprachen zu geschlossenen Sprachkenntnissen zu kommen. Vielmehr muss man sich das Ziel setzen, dass die Schüler motiviert und befähigt werden, ihre Sprachkenntnisse nach dem Schulabschluss zu entwickeln und zu vervollständigen. Ein realistisches Bild vom Verhältnis zwischen den Zielsetzungen und den zur Verfügung stehenden Möglichkeiten ist also besonders wichtig.

Im Hinblick auf die Gesundung der Sprachunterrichtsmethoden ist es sehr wertvoll gewesen, dass neuerdings auch in der Sprachdidaktik besonderer Nachdruck gelegt wird auf die Wichtigkeit der pragmatischen Gesichtspunkte (vgl. z.B. Kohonen 1983, Sajavaara 1984a, 1984b, Markkanen 1984, Lehtonen 1984). Dagegen war in dem traditionellen Fremdsprachenunterricht in der Tat häufig verstossen worden. Die Akzentverlagerung auf die aktuelle Sprachverwendung trägt eben zur Schaffung einer richtigen Motivation bei, weil bei den Schülern auch das Gefühl entstehen muss, dass sie in der Schule eine anwendbare Sprache erlernen.

Worauf die Schwerpunkte im einzelnen gelegt werden, hängt natürlich von der linguistischen Struktur der betreffenden Sprache ab. Bei einem flektierenden Sprachbau, wie das Deutsche ihn aufweist, muss in der Anfangsphase die Automatisierung ausgewählter, zentraler, kommunikativ relevanter Strukturen eine wichtige Rolle einnehmen. Wie eng die Kriterien Kommunikativität und Sprachrichtigkeit im Gegensatz zu dilettantischen Klischeevorstellungen miteinander

verflochten sein können, beweisen z.B. die Kasus- und Numerusformen der Substantive und Pronomina sowie die Personalformen der Verben im Deutschen (vgl. Luukkainen 1980). Die Fehler setzen den Sprecher nicht nur in ein ungünstiges Licht, sondern verhindern die Verständigung. Wenn also die an und für sich sehr interessanten ausersprachlichen "Interaktionskonventionen" auf Kosten der sprachlichen Gegebenheiten einseitig Übertrieben und auch in den Lehrmaterialien entsprechend gewichtet werden, kann ein Chaos entstehen. Das gilt selbstverständlich besonders für die Dritt- und Viertsprachen, wo es im Hinblick auf die Schaffung einer richtigen Motivation auch wichtig ist, den Schülern trotz beschränkter Lernzeit eine gewisse Sicherheit im Gebrauch der betreffenden Sprache zu vermitteln.

Wenn wir die obigen Grundsätze noch etwas konkretisieren wollen, können wir also sagen, dass im Rahmen des Unterrichts in Dritt- und Viertsprachen die Einfachheit und die Flüssigkeit als Ziel gelten. Das kommt auch darin zum Ausdruck, dass im Abitur ein Aufsatz der C- bzw. D-Sprache, der mit relativ einfachen Mitteln erzeugt ist und in dem die gewählten Ausdrücke und Strukturen richtig sitzen, höher bewertet werden kann als ein Aufsatz, der reich an komplizierten, aber mehr oder weniger falsch verwendeten Phrasen und Strukturen ist, die die Kommunikation beeinträchtigen. Wenn der Schüler seine Gedanken einfach, aber klar und flüssig ausdrücken kann, hat er das Lernziel vollauf erreicht. Dann ist sein Sprachgefühl derart entwickelt, dass er imstande ist, seine in der Schule erworbenen Sprachkenntnisse später zu bereichern.

Zu der Opposition Wissen/Können im Bereich der kommunikativen Kompetenz werde ich weiter unten noch explizit Stellung nehmen. Zuerst jedoch einige Worte zum Begriff Sprachgefühl, der in meinem Referat eine zentrale Stellung einnimmt und sich mit den beiden genannten oppositionellen Begriffen überschneidet. Bevor Sprachflüssigkeit erzielt werden kann, muss also Sprachgefühl vorhanden sein. Das letztgenannte setzt sowohl Wissen als auch Können voraus. Dabei darf das Wissen das Können nicht fesseln, sondern es muss das Können befestigen. Vor einiger Zeit kam mir das 1940 erschienene Buch

"Vom ABC zum Sprachkunstwerk" von W.E. Süskind in die Hand, dessen Untertitel "Eine Sprachlehre für Erwachsene" lautet. Als eine Art Motto in diesem Buch gilt der folgende Satz, der auch ausgezeichnet zum Thema dieses Symposions passt: "Dazu nun taugt nicht die Regel, sondern das Gefühl."

In der Vorrede seines Buches bringt Süskind die bekannte Geschichte vom Tausendfüßler, den "ein böser Nachbar fragte, mit welchem Fuss er eigentlich zu gehen beginne? - ob welcher Frage der Tausendfüßler die Unschuld seiner Unbewusstheit ganz und gar verlor und fortan überhaupt nicht mehr vom Flecke kam." Was ich hier unter Sprachgefühl verstehe, hat eben viel mit dieser "Unschuld der Unbewusstheit" gemein. Bevor man zu dieser Stufe des Könnens kommt, braucht man viel Übung, viel Automatisierung. Es ist dann wie beim Autofahren (vgl. Lampinen 1984) oder beim Volleyballspiel - sobald wir nicht bewusst an die Regeln denken, sondern uns auf das Fahren und Spielen konzentrieren, geht alles wie von selbst. Um kommunizieren zu können, müssen wir uns eben auf den Inhalt der Aussage konzentrieren, nicht auf ihre äussere Form.

Als Gegengewicht dazu, was ich oben zum Sprachrichtigkeitskriterium festgestellt habe, ist es wohl am Platze, noch das folgende Zitat von Süskind (S. 11) anzuführen:

"Sprachkunde ist keine regelsetzende, sondern eine lebensbeschreibende Wissenschaft, bei der man allein zu unterscheiden hat zwischen wirklich und unwirklich, zwischen lebendig und tot. Darum wird man in unserem Buch kaum je eingeteilt finden nach dem Schlüssel: Richtig ist falsch ist...., dafür aber immer die Mahnung vernehmen, die wir uns selber vorhalten: dass es das Lebendige zeigen gilt und nicht das Korrekte."

Bei der Diskussion dieser Probleme können wir mehrere oppositionelle Begriffspaare unterscheiden, die keineswegs Gegensätze darstellen, sondern sich überschneiden und im Sprachunterricht miteinander in Übereinstimmung gebracht werden sollen. Dazu gehören eben die Begriffe Wissen und Können, aber ausserdem auch Begriffe wie Lernfach und Übungsfach, kognitiver Aspekt und kommunikativer Aspekt, Grammatik und Lexik.

Die Relationen zwischen Wissen und Können bilden ein fundamentales sprachdidaktisches Problem, was auch darin zum Ausdruck kommt, dass wir diesem Thema bereits ein zweites Herbstsymposium widmen (vgl. Sajavaara & Leivo & Eloranta 1984). Weil sich die Begriffe Wissen und Können wegen ihrer gegenseitigen Verflechtung nicht in handgreiflicher Weise voneinander isolieren lassen, ist eine gewisse Unschlüssigkeit ihnen gegenüber verständlich. Meiner Ansicht nach kann der Unterschied so beschrieben werden, dass das Wissen eindeutig ein Bestandteil der linguistischen Kompetenz ist, die sich ihrerseits weitgehend mit dem Begriff Sprachgefühl deckt. Dagegen gehört das Können einerseits zur linguistischen Kompetenz bzw. zum Sprachgefühl, andererseits zur Performanz, d.h. zur aktuellen Sprachverwendung. Als reiner Begriff der Performanz bildet die Realisation den Gegenpol zum Wissen. Das verbindende Glied zwischen beiden ist also das Können. Darunter ist sowohl die Fähigkeit der linguistischen Organisation als auch die der situativen Verwendung zu verstehen.

Wenn wir also die vorhandenen Modelle der kommunikativen Kompetenz (vgl. z.B. Kohonen 1983, Wunderlich 1970) einerseits etwas reduzieren und andererseits etwas präzisieren oder mit anderen Worten denselben Vorgang unter einem anderen Aspekt beschreiben, ergibt sich die folgende Skizze:

Kommunikative Kompetenz:



Das Ziel im Fremdsprachenunterricht ist zweifelsohne das Können. Bevor aber Können vorhanden sein kann, braucht man Wissen. Wie aber das Wissen auf der jeweiligen Stufe und in der jeweiligen Sprache dargeboten wird, ohne dass es dem Schüler geht wie dem Tausendfüßler in der obigen Geschichte, bildet das primäre sprachdidaktische Problem. Eins steht fest: Zwischen den drei Gliedern Wissen, Können und Realisation besteht eine dauernde Wechselwirkung, was in der Skizze durch die zweiköpfigen Pfeile ausgedrückt wird. Je mehr die Sprache verwendet wird, desto fester wird das Können und desto höher wird die linguistische Kompetenz. Wir können also auch daraus schliessen, wie wichtig die Übung, die Automatisierung und die Darbietung von fremdem Sprachmaterial in allen möglichen Zusammenhängen im Sprachunterricht sind.

Die Bezeichnungen "Lernfach" und "Übungsfach", wie sie im finnischen Schulsystem seit eh und je vorkommen, stellen m. E. eine überaus künstliche und höchst undemokratische Gegenüberstellung dar. Einerseits muss jedes Fach in der Hierarchie der Schule ebenso wichtig sein, und zwar unabhängig davon, ob die geistigen oder die körperlichen Fähigkeiten trainiert werden (vgl. Harmia & Luukkainen 1977). Andererseits müssen gerade die Sprachen als "Übungsfächer" bzw. "Kunstfächer" gelten. Im Erwerb von Sprachen kommt es mehr auf die Übung als auf die Theorie an (vgl. Lampinen 1984).

Das gilt natürlich auch für die Dritt- und Viertsprachen. Nachdem der Lernstoff mit starker Hand reduziert worden ist, muss reichlich Zeit für das Einüben der verbliebenen Kapitel vorgesehen werden. Ich befürchte jedoch, dass auch in dieser Hinsicht die Reform der gymnasialen Oberstufe die Lernverhältnisse in der Schule verschlechtert hat. Infolge des Periodensystems werden gewisse Fächer zeitweise überhaupt nicht gelernt bzw. geübt. Aus stundenplantechnischen Gründen fällt dann der Unterricht in den Dritt- und Viertsprachen auch während der Lernperiode oft ungünstig aus, weil u.a. Doppelstunden vorliegen, die überhaupt nicht in den Unterricht von diesen Sprachen passen.

Auch die Begriffe Kognitivität und Kommunikativität hängen mit den Begriffen Wissen und Können eng zusammen. Ebenso schliessen sie sich nicht aus, sondern ergänzen sich. Wenn man von Grammatik spricht, schwebt einem vielleicht ein Bild zweckloser kognitiver Meditation vor. Oben wurde jedoch angedeutet, wie die Fehler in Strukturen die Kommunikation beeinträchtigen können. Die Verbindung zwischen Kognitivität und Kommunikativität kommt auch in der oben angeführten Skizze der kommunikativen Kompetenz zum Ausdruck. So gefährlich es auch sein kann, den Anteil der intellektuellen Analyse im Sprachunterricht zu übertreiben, müssen wir jedoch hier der Realität Rechnung tragen, dass die Schüler der gymnasialen Oberstufe auch selbst über Zusammenhänge nachdenken und für sie eine kognitive Erklärung erhalten wollen. Sie haben auch bereits Erfahrungen gesammelt über die Struktur anderer Sprachen.

Unter diesen Umständen beschleunigt und erleichtert eine gewisse kognitive Einstellung den Spracherwerb. Man darf aber nicht auf das leblose Memorieren der grammatischen Regeln zurückkehren, was typisch für die traditionelle Methode war, sondern muss die Wichtigkeit des angewandten Wissens im Auge behalten. Aus demselben Grunde befremden mich Verfahren, in denen sogar Schüler der Oberstufe aufgefordert werden, fremdsprachige Dialoge auswendig zu lernen. Auch das erinnert an den schulmeisterlichen Ton der alten Schule. Andererseits gilt eben hier, dass die Person des Lehrers entscheidet, welche Methode die Schüler zum besten Lernen motiviert.

In seinem bemerkenswerten Aufsatz "Ein gutes Gedächtnis kann effektivem Lernen hinderlich sein" behandelt Vuorinen (1984) verschiedene Lernverfahren unter psychologischem Aspekt. Er unterstreicht, dass das effektive Denken die wichtigste Bedingung für ein effektives Erlernen bildet, und dass der Denkvorgang gerade im Lernprozess selbst das Entscheidende ist und eine spätere kreative Anwendung des angeeigneten Wissens ermöglicht. Er schliesst den Beitrag mit dem Satz: "Die Zeiten des mechanischen Auswendiglernens sind endgültig vorbei - zum Glück der Lehrer und Schüler." Vuorinens Ausführungen über die Wichtigkeit des kognitiven Elements im

Fremdsprachenerwerb verdienen dringende Beachtung, und zwar bei der Motivation der Schüler und bei der Entwicklung des Sprachgefühls. Dieser Aspekt war auch entscheidend bei dem Lehrexperiment, das im Bereich der deutschen Sprache 1964-1976 in Töölön yhteiskoulu durchgeführt wurde und auf dessen Inhalt ich weiter unten noch einmal zurückkomme.

Zunächst müssen wir jedoch zu der Opposition Grammatik/Lexik Stellung nehmen. Nachdem die aktiv zu erlernenden Strukturen ausgewählt worden sind, müssen sie in einer sinnvollen Weise automatisiert werden, damit die Schüler die angestrebte Sicherheit und Flüssigkeit erlangen können. Wenn diese Automatisierung in der Schule nicht stattfindet, kann sie auch nach der Schule kaum erfolgen. Gleichzeitig muss kommunikativ relevante Lexik erlernt werden. In den Abituraufsätzen ist zu sehen, wie die falsche Wortwahl im Hinblick auf die Kommunikation oft gravierender wirkt als ein Strukturfehler. Doch ist die Erweiterung des Wortschatzes in der Hauptsache ein lebenslanger Prozess. Weil die Lexik im Gegensatz zur Grammatik aus Einzelfällen besteht, kann in der Schule nur der erste Grundstock gelegt werden. Derselbe Grundsatz ist auch in den von der Obersten Schulbehörde 1981 festgelegten Richtlinien für den Fremdsprachenunterricht auf der reformierten gymnasialen Oberstufe niedergeschrieben (S. 52 bzw. 54):

"Heute sieht man immer deutlicher, dass die syntaktische und morphologische Struktur der Sprache den Kern der Sprache bilden. Die flüssige Erzeugung der Sprache fordert eine automatische Beherrschung der Grundstrukturen. Im Rahmen der Lexik kann der Sprecher auch anhand weniger zentraler Strukturen eine schier unendliche Menge von neuen Sätzen erzeugen. Deshalb ist es in der Schule zweckmässig, Aufmerksamkeit ständig auf das Lehren von Strukturen zu richten."

"Die Lexik ist ein wichtiges Element im kommunikativen Sprachunterricht. Die Schüler sollen einen soliden Grundwortschatz erlernen und fähig werden, ihn flüssig zu gebrauchen. Das Erlernen von Lexik ist ein lebenslanger Prozess, aber schon während der Schulzeit muss eine Grundlage geschaffen werden, die das Erlernen des Wortschatzes auf neuen Lebensgebieten je nach Aufgabenbereich und Interessen ermöglicht."

Gerade im Umfang des Wortschatzes steckt m. E. der hauptsächlichste Unterschied zwischen den sog. A/B-Sprachen und den C/D-Sprachen. Bei der Lektüre einiger für die D-Sprache bestimmter Deutschbücher (ich meine vor allem das Lehrwerk "Ziel") gewinnt man den Eindruck, dass die Lesestücke von Anfang an mit reichlichem, kompliziertem Vokabular gefüllt sind, während meiner Ansicht nach die Einfachheit und Übersichtlichkeit in diesem Lehrgang das Ziel sein sollte. Derselbe Vorwurf kann wohl generell gegen alle Kurse der reformierten gymnasialen Oberstufe erhoben werden, die aus den Reaktionen der Lehrer und der Öffentlichkeit zu schliessen in der Regel zu sehr vollgestopft sind. Es ist schädlich im Hinblick auf die Motivation der Schüler, wenn eine grosse Kluft zwischen den Zielsetzungen und der Realität existiert.

Aus einigen ausländischen Untersuchungen geht hervor, dass der schlechte Erfolg im Fremdspracherwerb u.a. auf die mangelhafte Motivation und die schwache auditive Fähigkeit zurückzuführen ist (vgl. Pimsleur & Sundland & McIntyre 1964). In dem oben genannten Projekt in Töölön yhteiskoulu ging es vor allem darum, die Motivation der Schüler zu verbessern und zu einer günstigen Entwicklung des Sprachgefühls anhand reichlicher auditiver Übung beizutragen (Harmia & Luukkainen 1977). Im Hinblick auf die Motivation bildete das Bemühen den Ausgangspunkt, den Schülern das Gefühl zu vermitteln, dass sie die Sprache nicht als Selbstzweck, sondern als Mittel zum Zweck erlernen. Deshalb wurde besonderer Wert darauf gelegt, dass die Schüler von der ersten Stunde an das Zusammenspiel von Hören und Sprechen üben und von demselben Augenblick an in der betreffenden Sprache kommunizieren können bzw. dürfen, anstatt z.B. lange muttersprachliche Informationen über Landeskunde u.dgl. zu hören.

Wegen Mangel an Zeit muss m. E. nach wie vor in allen möglichen Situationen der Klasse die Zielsprache verwendet werden. Es ist unrealistisch, sich einzubilden, dass man in den Sprachstunden eine genügende Anzahl von authentischen Kommunikationssituationen simulieren könnte, um den Schülern die Vielzahl der Einzelfälle bei-

zubringen. Im Gegensatz zu den Vorstellungen während des goldenen Zeitalters der Direktmethode weiss man heute, dass der Fremdspracherwerb in der Schule grundsätzlich anders vor sich geht als das Erlernen der Muttersprache.

Sajavaara (1984a) behandelt ausführlich die Schwierigkeit des kommunikativen Sprachunterrichts in der normalen Klassensituation. Er kommt zu dem Schluss, dass es bei uns normalerweise keinen wirklichen kommunikativen Sprachunterricht, sondern vielmehr einen kommunikativen Grammatikunterricht gebe. Die Grammatik, nicht die Sprachverwendung werde kommunikativ gelehrt. Ausserdem werde die Kommunikativität häufig als Situationsbezogenheit missverstanden, wobei das Beibringen einiger schematischer Verhaltensmuster nur ein Zerrbild dessen darstelle, was in wirklichen Kommunikationssituationen stattfindet.

Auch angesichts dieser Realitäten scheint es mir nur richtig und ökonomisch zu sein, in den natürlichen Situationen der Klasse und - wenn es anders nicht möglich ist - in ganz "altmodischer" Weise in der normalen Klassensprache den Grundsatz "der vernünftigen Einsprachigkeit" anzuwenden, wobei die Zuflucht zur Muttersprache nur genommen wird, wenn sie den kürzesten Weg bietet. Dieses Prinzip schliesst auch nicht die Möglichkeit der Übersetzung oder die Berücksichtigung der kontrastiven Aspekte aus. Obwohl diese Methode "der vernünftigen Einsprachigkeit" nicht das von Sajavaara gewünschte Niveau der Kommunikativität erzielt, ist sie doch ein effektives Mittel bei der Entwicklung des Sprachgefühls der Schüler.

Vor diesem Hintergrund finde ich es ungünstig, dass die Lehrmaterialien der heutigen gymnasialen Oberstufe auf einer reichlichen Anwendung der Muttersprache aufzubauen scheinen. Als typisches Beispiel für diese regressive Entwicklung seien folgende Anweisungen in dem Lehrerheft von "Ziel" genannt (es handelt sich um das Beibringen der Begrüssung am Anfang des Lehrgangs): "Der Lehrer kann (zuerst sogar auf finnisch, dann auf deutsch) die Begrüssungssitua-

tion des Buches spielen, indem er selbst sowohl Günter als auch Karin ist." Meiner Ansicht nach wird hier der Schüler unterschätzt. Es ist schade, wenn die knappe Zeit mit muttersprachlicher Konversation vergeudet wird. Der Lehrer kann genausogut er selbst sein und die Klasse auf deutsch ansprechen.

Oben wurde bereits erwähnt, dass auditive Übungen zur Entwicklung des Sprachgefühls beitragen. Um das auditive Gedächtnis ("memory span") zu verbessern, empfiehlt es sich, den Schülern von Anfang an das aktive Hören anzugewöhnen, ohne dass schriftliche Stimuli vorliegen. Man kann von ganz einfachen Identifizierungs- und Wiederholungsaufgaben ausgehen und allmählich zu immer anspruchsvolleren Übungen übergehen, wo die Schüler auch lange Abschnitte im Gedächtnis behalten bzw. auf sie reagieren sollen. Die Gestaltungsfähigkeit macht ein wesentliches Element der Sprachkenntnisse aus, weshalb solche Übungen besonders wertvoll sind. Demselben Zweck dient die reichliche extensive Lektüre in der letzten Phase des Fremdspracherwerbs an der Schule. Mit Recht betonen die Richtlinien der Obersten Schulbehörde die Wichtigkeit der Entwicklung der Verständnisstrategien. Wie aus den Lehrbüchern der betreffenden Lehrgänge zu sehen ist, wird am Ende der Schule ein recht anspruchsvolles rezeptives Können vorausgesetzt, während die Anforderungen auf der produktiven Seite ziemlich bescheiden sind.

Ein wesentlicher Aspekt bei den Sprachkenntnissen ist jedoch auch die Kreativität, weshalb ihr auf die eine oder andere Weise Rechnung getragen werden sollte. Es gilt auch hier, dass man von Schülern nicht etwas verlangen darf, was nicht gelehrt worden ist, weil das Gefühl der Geborgenheit im Hinblick auf die Motivation wichtig ist. Zum effektiven Erlernen gehört, dass ein gewisser Themenkreis in Form von intensiver bzw. extensiver Lektüre und entsprechendem auditivem Material vorbereitet worden ist, bevor schriftliche oder mündliche Referate erwartet werden (vgl. dazu Harmia & Luukkainen 1977). Den Schülern müssen Themen gegeben werden, die sie besonders interessieren. Wenn man sie z.B. nur einige Sätze über eine Angelegenheit schreiben lässt, die sie gerade in dem Augenblick

fasziniert oder aufregt, wird man Überrascht sein, wie flüssig sie sich plötzlich mit ihrem bescheidenen Vokabular ausdrücken können! Sie konzentrieren sich auf den Inhalt, nicht auf die äussere Form und vergessen ihre Angst vor Fehlern.

Wie aus den Tabellen am Anfang dieses Referats hervorgeht, lernt die überwiegende Mehrheit der Schüler eben Deutsch als C- bzw. D-Sprache. Deutsch wird auch in der Regel von Jungen gewählt. Die Wahl von Französisch und Russisch dürfte auf einer ernsthaften Überlegung beruhen als die Wahl von Deutsch. Das müssen wir wohl berücksichtigen, wenn wir die Lernergebnisse in diesen Sprachen einschätzen und vergleichen. In unserer Untersuchung, wo die Sprachlernerfahrung und -motivation von finnischen und deutschen Studenten analysiert wurde, stellte sich als universaler Zug heraus, dass in beiden Ländern die Mädchen motivierter waren als die Jungen, andererseits aber die Mathematiker motivierter als die Neusprachler. Die Untersuchung ergab ferner, dass die finnischen Probanden die in der Schule erlernten Sprachen aufgrund der Motivation in folgende Reihenfolge brachten: Französisch, unmittelbar gefolgt von Englisch, dann Russisch und erst nach einem kleinen Abstand Deutsch, zuletzt Schwedisch. Unter den Schülern, die in der Schule Französisch gelernt hatten, gab es keinen so deutlichen Unterschied in der Motivation der Mädchen und Jungen wie in der Motivation der Schüler in anderen Sprachen (vgl. Luukkainen & Schröder 1981).

Die folgende Tabelle zeigt die Anzahl der Jungen und Mädchen, die im Frühjahr 1984 in einer C- bzw. D-Sprache an der Reifeprüfung teilgenommen haben sowie die Prozentzahlen für die höchste Note Laudatur bei beiden Geschlechtern:

	Insges.	Jungen	Mädchen	L a u d a t u r		
				Jungen	Mädchen	Insges.
Deutsch	14860	35,1%	64,9%	17,8%	23,0%	21,2%
Franz.	3113	13,7%	86,3%	25,0%	22,3%	22,7%
Russ.	1810	19,3%	80,7%	17,5%	22,0%	21,1%

Es stellt sich auch hier heraus, dass der Anteil der Jungen unter den Teilnehmern an der Prüfung in Deutsch weit grösser ist als unter den Teilnehmern an der Prüfung in Französisch und Russisch. Andererseits zeigt sich auch, dass der Anteil der höchsten Note Laudatur sowohl in Deutsch als auch in Russisch bei den Mädchen grösser ist als bei den Jungen. In Französisch dagegen verhält es sich umgekehrt, und die starke Motivation der Jungen, die diese seltenere Wahl getroffen haben, kommt auch hier zum Ausdruck. Das Verhältnis zwischen Motivation und Lernerfolg wird also auch in dieser Weise bestätigt. Andererseits zeigen die Statistiken des Abiturausschusses, dass die Abiturienten, die an der fakultativen Prüfung in Französisch teilnehmen, im allgemeinen auch in anderen Fächern etwas besser sind als die Abiturienten, die Deutsch und Russisch wählen, was sich dann in einer etwas höheren Quote für Laudatur-Noten äussert.

Zum Schluss wollen wir noch den Schwierigkeitsgrad der Prüfungen für Deutsch, Französisch und Russisch im Frühjahr 1984 miteinander vergleichen. Nach meinen eigenen Erfahrungen funktioniert ein Abiturtest zuverlässig, wenn die Anzahl der richtigen Lösungen etwa $\frac{2}{3}$, d.h. 60-70%, des Maximums beträgt (vgl. Luukkainen 1979). Wenn er dagegen wesentlich schwieriger oder wesentlich leichter ist, verliert er an Relevanz. In der folgenden Tabelle werden die Zahlen für richtige Lösungen im Hör- und Textverständnistest sowie im Strukturtest angegeben; ausserdem wird die durchschnittliche Note im Aufsatz hinzugefügt. Es ist zu beachten, dass hier jeweils Prozentwerte, nicht absolute Punkte stehen (die jeweiligen maximalen absoluten Punkte betragen bekanntlich 90, 90, 20 und 99). Bei den Hör- und Textverständnistests wird auch die Reliabilität in Klammern angegeben.

	Hörverst.	Textverst.	Strukt.	Aufsatz
Deutsch	69,2% (.81)	69,7% (.84)	64,6%	72,9%
Franz.	75,4% (.78)	74,1% (.81)	70,2%	74,7%
Russ.	81,3% (.75)	83,3% (.67)	72,0%	72,3%

Obwohl die Reliabilität eines Tests von vielen verschiedenen Faktoren abhängt, kann man auch hier sehen, dass mit abnehmendem Schwierigkeitsgrad auch die Reliabilität oft sinkt. Im ganzen ist die Ausarbeitung solcher Tests eine anspruchsvolle und viel Sachkenntnis verlangende Aufgabe - vor allem weil es keine Möglichkeit gibt, ihren Schwierigkeitsgrad und ihre Zuverlässigkeit vorher zu testen.

Die Form der Abschlussprüfung ist ein ausserordentlich belangvolles sprachdidaktisches Problem, weil sie die Arbeit der Schule prägt, ob wir es wollen oder nicht. Es gehört vielleicht zur Natur einer landesweiten Prüfung, dass sie etwas konservativer Art ist und nicht ohne weiteres jeder Moderichtung folgen kann. Eins steht jedoch fest: Auch wenn ein mit Sorgfalt und Sachkenntnis ausgearbeitetes Multiple-choice-Verfahren zuverlässig funktionieren kann, ist in der jetzigen Prüfung jedoch zuviel davon enthalten.

Bibliographie

Harmia, S. & M. Luukkainen, 1977. "Mitä, miten ja miksi opetamme. Kokemuksia saksan opetuksesta Töölön yhteiskoulussa vuosina 1964-1976." Kasvatus, vol. 4/8: 225-231.

Huju, L. & R. Markus & P. Landgraf & I. Nuotio & M. Sinisalo, 1982a. Ziel I. Texte 1-3. Vaasa: Kirjayhtymä.

Huju, L. & R. Markus & P. Landgraf & I. Nuotio & M. Sinisalo, 1982b. Ziel I. Opettajan opas. Helsinki: Kirjayhtymä.

Kohonen, V., 1983. "Learners, teachers and graded objectives in communicative language teaching". In Nyyssönen & Mauranen (ed.): 37-67.

Lampinen, P., 1984. "Kielitieto ja kielitaito vieraan kielen luokassa". In Sajavaara & Leiwo & Eloranta (ed.): 41-44.

Lehtonen, J., 1984. "Vaikeneva kulttuuri ja vieraan kielen pelko". In Sajavaara & Leiwo & Eloranta (ed.): 75-94.

Luukkainen, M., 1979. "Tekstinymmärtämisen mittaamisesta ylioppilastutkinnoissa keväällä 1977 ja 1978". Tempus, vol. 3/14: 12-19.

Luukkainen, M., 1980. "Miten kielitaito heijastuu kirjoittelussa". *Tempus*, vol. 9/15: 8-11.

Luukkainen, M. & K. Schröder, 1981. Diversifikation und Motivation im Fremdsprachenunterricht. Eine vergleichende Studie zwischen Finnland und der Bundesrepublik Deutschland. Publications de l'association Finlandaise de linguistique appliquée 29. Turku.

Markkanen, R., 1984. "Puheaktitutkimus ja kielenopetus". In Sajavaara & Leiwo & Eloranta (ed.): 57-63.

Nyyssönen, H. & A. Mauranen (ed.), 1983. *AFinLA Yearbook 1983*. Publications de l'association Finlandaise de linguistique appliquée 36. Oulu/Helsinki.

Pimsleur, P. & D. Sundland & R. McIntyre, 1964. "Under-Achievement in Foreign Languages". *Iral*, vol. 2: 113-150.

Richtlinien für den Fremdsprachenunterricht, s. Vieraiden kielen opetus kurssimuotoisessa lukiossa.

Sajavaara, K., 1984a. "Kielitieto ja kielitaito kieliluokassa". In Sajavaara & Leiwo & Eloranta (ed.): 65-73.

Sajavaara, K., 1984b. "Kieliooppi ja kommunikatiivinen kielenopetus". *Tempus*, vol. 9/19: 8-9.

Sajavaara, K. & M. Leiwo & M. Eloranta (ed.), 1984. *AFinLA Yearbook 1984*. Publications de l'association Finlandaise de linguistique appliquée 38. Jyväskylä.

Süskind, W.E., 1940. Vom ABC zum Sprachkunstwerk. Eine deutsche Sprachlehre für Erwachsene. Stuttgart: Deutsche Verlags-Anstalt.

Vieraiden kielen opetus kurssimuotoisessa lukiossa, 1981. Kouluhallitus.

Vuorinen, R., 1984. "Hyvä muisti saattaa olla tehokkaan opiskelun esteenä". *Tempus*, vol. 8/19: 8-10.

Wunderlich, D., 1970. "Die Rolle der Pragmatik in der Linguistik". *Der Deutschunterricht*, vol. 4/22: 5-41.

OM SPRÅKLIGA DRAG OCH FELTYPER VID FRI SKRIFTLIG FRAMSTÄLLNING I FRÄMMANDE SPRÅK

Antti J. Pitkänen (Tammerfors universitet)

1 Bakgrund och syfte

I detta föredrag presenteras ett antal preliminära resultat från projektet Abiturientsvenska, som fick sin verksamhetsstart år 1980 på Institutionen för filologi II vid Tammerfors universitet. Projektets målsättning är att beskriva de språkdrag och feltyper som är kännetecknande för finskspråkiga abiturenters svenskuppsatser i studentexamen. Resultaten är utan tvivel av intresse också med tanke på andra språklännen. Därför har här och där paralleller dragits i synnerhet med tyskan.

Redovisningen bygger på en förstudie med 250 uppsatser från studentexamen i svenska språket åren 1978-80. Temat skrivförmåga/fri skriftlig framställning har behandlats på metodkurser och seminarier inom ämnet nordisk filologi. Arbetet har redan avkastat ett halvtduusin pro gradu -uppsatser (bl.a. Jokiranta 1978, Nyysti 1983, Hirvikoski 1983, Luotonen 1985). Nämnas bör också temanumren Interlingua-studier i Institutionens skriftserie (nr 7 och 10). Detta bidrag publiceras i informationsyfte med tanke på fortbildning för språklärare. Förbehåll måste göras för ändringar speciellt av kvantitativa uppgifter, som kommer att justeras i den slutliga resultatredovisningen.

2 Några performansdrag

Medellängden för svenskuppsatserna i studentexamen är ca 200 löpord. Flickorna skriver något längre uppsatser än pojkarna. Långa uppsatser tenderar att ge ett bättre betyg än korta. Enligt Studentexamenenämndens (SEN:s) anvisningar är ideallängden 150 - 200 ord.

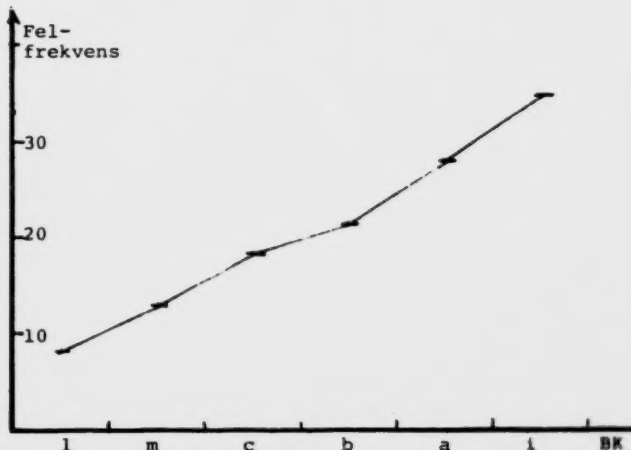
Lexikalt sett utmärks goda uppsatser av en stor andel lågfrekventa ord och engångsord. Goda skribenter visar en bättre förmåga att variera ordförrådet. De använder färre hjälpverb, men flera egentliga huvudverb än svaga skribenter. Könskillnaderna är minimala. (Jfr Meriläinen 1982.)

Syntaxen i uppsatserna motsvarar väl det inlärdas standardschemat för sats och mening. Meningslängden tycks inte ha direkt samband med betyget. Flickorna skriver något kortare satser än pojkarna. En uppsats omfattar i genomsnitt 16 - 17 huvudsatser och 10 - 11 bisatser. Andelen bisatser är oväntat hög, särskilt i svaga uppsatser. En mindre variation av bisatstyper och ett stort överskott av att-satser (typen: ... påstår/tror/anser/tycker att ...) tyder på en sämre förmåga till meningsplanering hos svaga skribenter. Grafiska meningar (textsekvenser mellan "stora" skiljetecken) sammanfaller i regel med makrosyntagmer (ungefär: huvudsats och dess underordnade satser och satsmotsvarigheter). Påståendemakrosyntagmerna dominerar till 95 procent. Frågesatser används i någon mån. Meningsfragment är rariteter. Variationen av meningstyp är sålunda minimal och förekommer nämnvärt endast i de bästa svenskuppsatserna.

3 Felfrekvens och betygskategori

Det är inte helt oväntat att felfrekvensen korrelerar med betygskategori. Redan Hultman & Westman 1977 har lagt märke till detta vid analys av gymnasiestuppsatserna i modersmålet svenska. Överraskande är dock att korrelationen tycks vara nästan lineär (se fig. 1 nedan). Resultatet är så mycket mer överraskande som en stor del av svensklärarna tycks anse att de bedömer uppsatser som helhet och inte fäster sig vid fel. Det är uppenbart att språkets korrekthet/uppsatsernas höga felhalt i alla fall har samband med de egenskaper som gör en uppsats "klar och lättläst"/"svårsläst", "dunkel" (jfr SEN:s kriterietabell och Hultman & Westman 1977: 231). - En fortfarande outredd fråga är vilken roll felens art spelar vid bedömningen och hur relevant våra språklärares bedömning är exempelvis med tanke på resp. språkstrukturs kommunikationsvärde.

Figur 1. Felfrekvens och betygskategori; felfrekvensen är beräknad för standardlängden 200 löpord. BK enl. SEN:s bedömning. Jfr Luotonen 1985: 35-.



Grovindelas felen i stav-, böjnings-, syntax- och ordfel kan felhalten i dessa huvudkategorier beräknas i betygskategorierna. SEN:s något vaga formuleringar i kriterietablen kan sålunda specificeras tentativt som följer (tab. 1).

Tabell 1. En uppsats på 200 ord innehåller i genomsnitt följande antal stav-, böjnings-, syntax- och ordfel i betygskategorierna I ... I. I kolumnen längst till höger ges ett förslag till felens variationsvidd vid bedömningen.

BK	stav-fel	böjnings-fel	syntax-fel	ord-fel	Förslag till variationsvidd
I	1	1	3	2	1 - 8
m	1-2	2	5-6	2-3	9 - 14
c	2-3	3	8	3	15 - 19
b	3	3-4	10	4	20 - 24
a	3	6	13	4-5	25 - 29
i	4	7	16	5-6	30 -

I belysning av feldistributionen i pilotmaterialet ryms en uppsats med betyget "c" sålunda i genomsnitt 2 - 3 ortografiska, tre morfologiska, hela åtta syntaktiska och tre ordsemantiska fel. Den föreslagna variationsvidden för betyget "c" är då 15 - 19 fel. Märkas bör dock att felens grovlek inte har beaktats i denna "grammatiska" kriterietabla. Övriga kategorier nog tolereras egentligen rätt många fel vid högsta betyget; felen är dock givetvis lindrigare än i lägre betygskategorier. En uppsats kan innehålla bortåt 30 fel (= "rikligt med språkfel"; SEN:s kriterietabla) och ändå bli godkänd. Också i betygskategorin "c" är det säkerligen fråga om relativt lindriga fel, som inte stör kommunikationen (SEN:s kriterietabla).

4 Om feltyperna

I kategoriindelningen ovan räknas felen när det gäller substantivets böjning, pron. (+ adj.) + subst. och verbets konjugation till böjningsfel. Species-, kongruens- och syftningsfel hör till (morfo)syntaxen, prepositionsfel till syntaxen. Ordfel är lexikalt-semantiska avvikelser från språknormen. Abiturienternas interimspråk (interlanguage; Selinker 1972) kännetecknas av ett stort antal syntaktiska och morfosyntaktiska fel. Ungefär vartannat fel utgörs av en avvikelse från den syntaktiska normen. Böjningsfel som ännu är typiska för uppsatser av eleverna i åk 1 på gymnasiet har minskat till 20 procent i abiturientuppsatserna. Lika vanliga är också ordfel. Ortografiska fel som är utmärkande för nybörjarstadiet, är relativt sett oväntat högfrekventa också i de högsta betygskategorierna (något över tio procent av samtliga fel). Pojkarna gör sig skyldiga till stav- och böjningsfel något oftare än flickorna. Språkformen i flickuppsatserna är snyggare än i pojkuppsatserna, eftersom felfrekvensen är något lägre.

Stavfelen är sålunda svåra att eliminera. En del av dem beror på interferens från moderamålet (typerna *nyförtiden, *akatemisk, *priat på personal). Ortografiska fel som kan förklaras med negativ transfer från de övriga främmande språken håller på att öka (t.ex. *job, *product, *so, *Finnland). Bl.a. inriktningen på muntlig språkträning gör att stavfelen inte kommer att minska i skriftlig framställning.

Till följd av en "uppluckning" vid ordinlärning kommer lärarna i flexionspråk som svenska och tyska att ha mycket att korrigera också i framtiden. Genus och deklination borde varje språkinlärare inpränta. Fel som *problemer (: problem), *faktor (: fakta) är sedan länge klassis-

ka i svenskuppsatserna. Allt elementärare fel dyker dock upp i abiturientuppsatserna. Den utbyggda nominalfrasens (pron. + adj. + subst.) böjning bereder stora svårigheter för våra gymnasister. Verbböjningen drabbas av grova övergeneraliseringar (typen: *hållde (: håll)).

Som anförts ovan gäller ungefär hälften av felen i abiturientuppsatserna syntaxen. Rätt ofta bryter skribenterna mot den nominala morfosyntaxen: en tredjedel av felen i denna kategori gäller nämligen kongruens/syftning resp. species. Vid kongruensfel är det oftast fråga om adjektivkongruens, oftast i numerus (typen: *reaktionerna är långsam (: långsamma)), men också i genus (typerna: *Det är säker (: säkert), *han är säkert (: säker)). Vid pronomenkongruens gäller felen ofta pron. själv (typen: *vi måste själv (: själva) fatta beslutet). Syftningsfel gäller också ofta numerus (typen: polisen - *de (: den), dessa frågor - *den (: de)). Kongruens- och syftningsfel är oväntat "seglivade". De är frekventa i uppsatserna utom i den högsta betygs kategorin. Olika typer av kongruens och syftning borde uppenbarligen tränas intensivare i svenska (och tyska) där de hör till språkets basstrukturer.

Speciesfelen är betydligt svårare att råda bot på. Speciessystemet i finskan är knappast fullkomligt kartlagt än. Den traditionella skolgrammatikens speciesregler bygger på givenheten och tycks inte fungera tillfredsställande i praktiken. Än så länge känner man bara en del fel-taxonomier: 8 - 9 av tio speciesfel utgörs av någon av följande feltyper:

- obestämd form i stället för bestämd form
(typen: *två gånger i en vecka (: i veckan, per vecka));
- bestämd form i stället för obestämd form
(typen: *det finns flickorna (: flickor) som inte gillar mig);

- överflödlig obestämd artikel

(typen: *han vill bli en redaktör (: redaktör)).

Därtill vet man att speciesfelen inte minskar utan ökar relativt sett i högre betygs kategorier. Dessutom har det visat sig att först två års vistelse i målspråkmiljö medför en väsentlig minskning av speciesfel (Pitkänen 1980). Endast med hjälp av longitudinella undersökningar skulle man kunna bevisa en eventuell förekomst av ett provisoriskt regelsystem för speciesbruket. Då nominalfrasen är en ytterst frekvent kategori i språket försämrar svårigheterna med species lätt det allmänna intryck som läsaren får av en uppsats.

Förutom speciesfel har även uppsatser två stora underavdelningar av syntaktiska fel: ledföljdsfel och prepositionsfel. Vart fjärde syntaxfel berör ledföljden i huvud- eller bisats. Vid vart fjärde ledföljdsfel rör det sig om placeringen av satsadverbial i bisatser (typen: *en plats som man kan alltid (: alltid kan) besöka). Nästan lika frekventa är felen i adverbialinledda påståendesatser (typen: *Nuförtiden man talar (: talar man) ofta om våldet i tv). Den s.k. eftersatsen är ett specialfall av föregående satstyp (*När han kom till oss, vi gick (: gick vi) tillsammans ut på stan). Till följd av allt starkare transfer från engelskan har följande feltyp blivit vanligare och vanligare: *han naturligtvis glömde (: han glömde naturligtvis) kvar paraplyet i bussen. Rätt många skribenter har svårigheter med att inleda en bisats (t.ex. med distinktionerna när/då - jfr. ty. wenn/als/weil, da - innan/förrän och med dubbel bisatsinledare: att när). Dessa felkällor är nog välkända, men lärarnas ansträngningar för att utrota dem tycks vara dömda att misslyckas - åtminstone i de lägre betygs kategorierna. Dåliga uppsatser vimlar nämligen av elementära ledföljdsfel. I de högre BK:na uppvägs de av mera komplicerade feltyper, så

att den relativa frekvensen för ledföljdsfel förblir rätt stabil genom hela betygsskalan.

I likhet med speciesfelen ökar även prepositionsfe-
- totalt en fjärdedel av syntaxfelen - relativt sett i de
högsta BK:na. Regelsystemet för prepositionsbruket är rikt
och finindelat i de germanska språken inom denna gränsszon
mellan grammatik och lexikon. Språklärarna kan dock med
rätt enkla medel bibringa eleverna effektiva förvaraneka-
niamer mot prepositionsfel. Då bruket av prepositioner
ofta bestäms av satssemantiska faktorer lönar det sig att
fästa uppmärksamhet vid fylligare kontexter i stället för
att låta eleverna plugga löeryckta prepositionsfraser.
Över hälften av prepositionsfe- len gäller nämligen rektion-
nen hos verb eller adjektiv. För det första tycks skriben-
terna ha svårigheter med rektionstyp: med eller utan pre-
position (typerna: *man ska tänka sina medmänniskor (:
tänka på), *Han frågade från mig (: frågade mig) om råd).
För det andra väljer skribenterna ofta fel preposition vid
ett prepositionsverb eller -adjektiv (typen: *njuta om (:
av) livet, *vara tacksam över (: för) något). Denna fel-
typ utgör minst en tredjedel av alla prepositionsfel. - Fel-
valda prepositioner vid rums- eller/och tidsuttryck utgör
"endast" en femtedel av prepositionsfe- len.

Prepositionsbruket i abiturienternas interimspråk
regleras fortfarande av ett provisoriskt regelsystem som i
hög grad kommit till genom förenklingsstrategier från
käll- och målspråket. Finskans allativ eller illativ mot-
svaras exempelvis ofta av prep. till i abiturientensvenskan
(typen: *ägna sig till (: åt) något; *detta är viktigt
till (: för) mig; *jag byter ut vanliga TV-program till (:
mot) Sky Channel). Prep. från eller av får motsvara fin-
skans elativ (typen: *jag har läst från/av (: om) honom i
en tidning; *bero av (: på) något). En del av dessa inte-
rimregler härrör uppenbarligen från målspråket (typen:

tycka om något -> *njute om (: av) något), i vissa fall även från andra språk, oftast engelska (typen: *vara intresserad i (: av) frilidrott <- eng. to be interested in athletics). Någon gång tycks flera faktorer motverka den korrekta inläringen (typen: fi. keksi viikkoa sitten, eng. two weeks ago -> *två veckor sedan (: för - sedan)).

Prepositionsfel kan i många fall motarbetas med hjälp av semantiska tumregler. Finskans inessiv (-asA) motsvaras i regel av prep. i (i Helsingfors, i Täby). I fråga om öar, slott, offentliga lokaler och tillställningar används dock prep. på (på Gotland, på Sveaborg, på en restaurang, på bio, på bröllop; jfr också ty. in Wartburg/auf der Wartburg, auf der Post). Ytterligare ett exempel på dylika tumregler: sinnesstämning uttrycks ofta med hjälp av prep. över i svenskan (och prep. über i tyskan): sörja över ngt/över etw. sorgen, sorgsen över ngt/över etw. traurig (undantag: nöjd, belåten med ngt/mit etw. zufrieden). Denna undervisningsstrategi är nog välkänd, men samlar knoppologi.

5 Slutord

Varje språklärare ställer sig väl positiv till en granskning av performansdrag, egenskaper som är typiska för ett interimspråk. En ingående redovisning av feltyper kan däremot oroa en vanlig lärare på fältet: "Jag bedömer ju inte enskilda fel! Jag undervisar ju egentligen inte i grammatik - och det är väl inte meningen heller!" Felens betydelse bör ingalunda överdrivas. Obestriddligen berättar de dock en hel del om hur undervisningen har anammats av eleverna och vilka svårigheter eleverna har. De berättar också indirekt om inlärningsprocessen och om reglerna för elevernas interimspråk.

Det är också rätt att inte överbetona vikten av grammatik och språkstrukturer i undervisningen (även om de är viktiga). Men skall egentligen inte dra en skarp gräns mellan grammatik och ordförråd. Inlärnigen sker bäst med stöd av meningsfulla (situationella) kontexter. Kollokationer (t.ex. i hög grad/i stor utsträckning) och flerledade idiom (t.ex. ta reda på ngt) spelar en central roll (även) i uppsatsskrivning. Det är exempelvis nyttigt att veta att finskans verb tarkoittaa motsvaras av sv. mena (ty. meinen) om det har ett mänskligt (+Mänsk) subjekt, eljest av sv. betyda, innebära (ty. bedeuten). Eller att finskans verb esittää har minst ett halvt dussin motsvarigheter i svenskan och tyskan, och att valet beror på de semantiska egenskaperna hos subjekt/objekt. En uppgift för språkläraren är att bibringa eleverna en förmåga att hantera dessa språkets byggklossar. Det är helt förkastligt att ge ett mekaniskt råd som: "strö in ett par, tre flerledade idiom (exv. ta hänsyn till, passa på, gå i land med), det gör susen! Språket ser finare ut och - ni får ett högre betyg!" En annan minst lika viktig uppgift är att ge en helhetsyn på en uppsats som en välformad text (i en given situation).

Referenser

Elevavenska, red. av Kent Larsson. Ord och stil, Språkvårdssamfundets skrifter 12. Studentlitteratur: Lund 1981.

Errata, red. av Jan Svartvik. G.W. Gleerup: Lund 1973.

Error analysis, red. av Jack C. Richards. Third impression. Longman: London 1977.

Hirvikoski, Hannele, 1983. De viktigaste feltyperna i bruket av prepositioner i finska abiturienters studentuppsatser. Maskinskriven bilsudatursavhandling. Tammerfors universitet.

Hultman, Tor G. & Westman, Margareta, 1977. Gymnasistsvenska. LiberLäromedel: Lund.

Interlingua-studier I-II, red. av Antti J. Pitkänen. Skrifter utg. av Institutionen för filologi II, nordisk filologi nr 7 och 10. Tammerfors 1982, 1984.

Jokiranta, Ritva, 1981. Syntaktiska studier över finska abiturienters studentuppsatser i svenska våren 1978. Maskinskriven pro gradu -avhandling. Tammerfors universitet.

Larsson, Kent, 1984. Skrivförmåga. Liber: Lund.

Luotonen, Ann-Mari, 1985. Om syntax och syntaktiska fel i finska abiturienters uppsatser i andra inhemska språket våren 1978-1980. Maskinskriven pro gradu -avhandling. Tammerfors universitet.

Meriläinen, Heikki, 1982, 1984. Beskrivning av språket i ett urval abiturientuppsatser från år 1980. 1-2. Interlingua-studier I, 57-111, Interlingua-studier II, 1-92.

Nyyti, Tuula, 1984. Om ordföljden i finska abiturienters studentuppsatser i svenska våren 1978. Maskinskriven pro gradu -avhandling. Tammerfors universitet.

Pettersson, Åke, 1982. Hur gymnasister skriver. Svensk-lärarserien 184. Trelleborg.

Pitkänen, Antti J. 1980. Vad språkfelen berättar. SPRINS-projektet 5. Göteborgs universitet. Göteborg.

-----, 1982a. Interlingua-studier - teori och praktik. Interlingua-studier 1, 1-35.

-----, 1982b. Om feltyper hos finsktalande elever med svenska som andra/främmande språk. Svenskans beskrivning 13, 277-288.

Selinker, Larry, 1972. Interlanguage. IRAL 3/X; myös teoksesta Error analysis, 31-54.

Studentexamenenämndens cirkulär 1/1980.

Svenskans beskrivning 12, red. av Mirja Saari och Marika Tandefelt. Meddelanden från Institutionen för nordiska språk och nordisk litteratur vid Helsingfors universitet B:6. Helsingfors 1982.

Teleman, Ulf, 1979. Språkrätt. LiberLäromedel: Lund.

TEKSTILAJIT JA KIRJOITETUN TEKSTIN MUOTOUTUMINEN

Pauli Roinila (Joensuun yliopisto)

Pidän lähtökohtanani kääntämisentutkimusta. Tämä lähtökohta on mielestäni siinä suhteessa otollinen, että kääntäminen on kielitaidon pitkälle menevä sovellus: onhan kääntäjä viime vaiheessa kirjoitetun tekstin tuottaja. Kirjoitetun tekstin muotoutumisprosessi on erikoispiirteistään huolimatta pitkälti samansuuntainen kuin kielenkäyttäjällä yleensä. Oleellista tekstilajien kannalta on se, että kääntämisen tutkimus on viime aikoina osoittanut vilkasta kiinnostusta tekstiluokituksia kohtaan ja useat sen edustajat näkevät tekstilajeissa käännöksen tärkeän ekvivalenssikriteerin.

Kääntämisestä lähteminen tarjoaa minulle mahdollisuuden yrittää samanaikaisesti palvella kahta herraa. Normaalisti suositeltu kääntämisen suunta vieraasta kielestä äidinkielen antaa tilaisuuden pohtia äidinkielen kannalta tekstilajeja ja kirjoitetun tekstin muotoutumista, käänteinen suunta puolestaan vie ajatukset vieraan kielen opetukseen.

Tekstiluokittelusta ei todellakaan ole puutetta. Omasta lähtökohdastani pidän varsin käyttökelpoisena Katharina Reissin luokittelua, joka lähtee tekstilajeista, erottaa myös läheisesti yhteenkuuluvien tekstilajien muodostamia tekstilajirympäleitä (Textklassensorte) ja abstrahoi sitten Bühlerin Organon-malliin nojautuen kielen funktioitten perusteella neljä tekstityyppiä: informatiivisen, ekspressiivisen, operatiivisen ja yhdistelmätyyppinä multimediaalisen tekstityypin. Tekstityyppi-kategoriasta tyydyn toteamaan sen, että sille ovat tyypillisiä tietynlainen universaalisuus ja sekatyypien runsas esiintyminen. Reiss muutoin huomauttaa, että kaunokirjallisten tekstien luokitteluun tekstityypijako soveltuu tekstilajijakoa paremmin.

Tekstilajin määrittelyssä Reiss nojautuu Friedemann Luxiin, jonka mukaan tekstilaji on "yliyksilöllinen puhe- ja

kirjoitusaktityyppi, joka liittyy toistuviin viestintätapah-
tumiin ja jolle samantapaisissa viestintäyhteyksissä käytet-
tyinä on kehittynyt tyypillisiä kieli- ja tekstimalleja".
Kun tekstilajeja näyttäisi karakterisoivan se, että ne ovat
kulttuurisidonnaisia ja että niihin vaikuttavat voimakkaasti
konventiot, on näitä konventioitakin pyritty kuvaamaan. Kon-
ventioiden piirteiksi on nähty niiden muuttuvuus kyseisessä
kulttuurissa, niiden joustavuus - ne kun eivät sido kielen-
käyttäjää läheskään yhtä paljon kuin esim. kieliopin säännöt
eivätkä näin ollen sulje pois idiolektisia piirteitä - sekä
niiden liittyminen ihmisen sosiaalistumisprosessiin.

Tekstilajien merkityksestä ollaan monta mieltä. Mainit-
semani Lux on sitä mieltä, että tekstilaji on osa tekstin
identiteettiä. Tekstin yksilöllisiä piirteitä, tekstiä teks-
tiyksilönä, tuskin kukaan haluaa kieltää, mutta tämän yksi-
öllisyyden korostamisessa on selviä eroja. Ollaanpa teksti-
lajien merkityksestä mitä mieltä tahansa, näytämme ainakin
tunnistavan niitä melko hyvin. Kuten oheisesta liitteestä
havaitaan, tässä Matthias Dimterin käyttämässä testissä, jos-
sa hänellä oli koehenkilönä eri alojen opiskelijoita, on kym-
menen tekstiä. Tuloksissa kiinnitetään huomio paitsi varsi-
naiseen valintaan, myös käytettyyn aikaan ja valinnan perus-
teluihin. Teksteissä on selväkielistä ainoastaan tietyt usein
toistuvat verbit (sein, haben, werden), artikkelit, pronomi-
nit, konjunktiot ja lukusanat. Jos meillä olisi enemmän aikaa
tekstien tarkasteluun, teksteihin sisältyvät tunnistamissig-
naalit alkaisivat varmasti viestittää. Työdyn tässä vain to-
teamaan, että liitteen tekstilajit ovat seuraavat:

1. sätiedotus
2. keitto-ohje
3. käyttöohje
4. seurapelin ohje
5. uutinen
6. sopimus
7. kuolinilmoitus
8. testamentti
9. avioliittoon halutaan -ilmoitus (Heiratsanzeige)
10. kommentti

Olen pysytellyt toistaiseksi enimmäkseen aiheotsikon vasemmalla puolella. Ennen kuin siirryn tekstilajien merkityksen ja käyttökelpoisuuden pohtimiseen, on paikallaan korostaa, että tekstilajitutkimus on vasta alussa. Sama pätee myös kontrastiiviseen tekstilajitutkimukseen, siis kahden kulttuurin tekstilajikäytäntöjen vertailuun. Esimerkiksi englannin ja saksan kielten välisistä suhteista tässä katsannossa on olemassa vasta muutamia suppeita tutkimuksia. Kuitenkin juuri kontrastiivinen tekstilajitutkimus voisi tuoda selvyyttä kysymykseen, mikä merkitys tekstilajikompetenssilla on esimerkiksi vieraskielisen kirjoitetun tekstin muotoutumisessa.

Jos tekstintuottajan kompetenssi on pitkälti myös tekstilajikompetenssia, siitä ilmeisesti seuraa, että kirjoittajan on hallittava useiden tekstilajien konventiot alkaen niiden tunnistamisesta ja päätyn niiden tuottamiseen. Tieto- ja taitokomponentit näyttäisivät tässä asiassa olevan lähes erottamattomassa liitossa. Samasta oletuksesta johtuisi edelleen, että tekstilajien hallintaa tulisi tietoisesti pyrkiä edistämään. Ilman muuta on selvää, että tällainen harjoitus pakostakin voi olla vain esimerkkeihin pohjautuvaa. Suurimman osan siitä voisivat kaiketi muodostaa normaalit opetustilanteet, joissa tekstilajeihin kiinnitettäisiin huomiota. Tekstilajeihin syventyminen vie ainakin niin äidinkielellä kuin vieraalla kielelläkin tapahtuvan kirjoittamisen keskelle elämää.

Haluan vielä kerran viitata kääntämisentutkimukseen haluamatta ollenkaan väittää, että sen piirissä viime aikoina korostettu tekstin viestinnällisen tarkoituksen esiin tuominen olisi mitenkään ainutlaatuista. Mutta samoin kuin kääntäjälle kohdetekstin muotoutumisprosessissa käännöksen funktio on aivan keskeinen ohjenuora, samoin on mielestäni tekstilajienkin rooli nähtävä kulloisenkin tehtävänannon valossa. Tuotettavan tekstin tarkoituksen ja sillä tavoiteltavan kohderyhmän hämäräksi jääminen saattaa johtaa siihen, että syn-

tyvät tekstit ovat eräänlaisia näennäistekstejä, joiden ääri-
viivattomuus on peräisin jo tekstin syntymistilanteesta.

Joudun lopuksi sivuamaan melko luettelonomaisesti eräi-
tä assosiaatioita, joita aihe herättää. Ja ihan rehellisesti
on sanottava, että mieluummin heittää ilmaan kysymysmerkkejä
kuin yrittää vastata kysymyksiin. Onko ajateltava, että teks-
tilajikonventioitten merkitys on siinä, että ne antavat kir-
joitetulle tekstille hahmoa, laukaisevat kohdekulttuurin vas-
taanottajassa tunnistamissignaaleja, ohjaavat hänen lukuodo-
tuksiaan, samoin kuin tuotetun tekstin ymmärtämistä? Voidaan-
ko tekstilajipiirteitä tai niiden puuttumista tekstin funktio
huomioon ottaen käyttää menestyksellisesti tuotoksen arvioin-
nin kriteereinä?

Palaan lyhyesti kysymykseen tekstien yksilöllisistä piir-
teistä. Eri tekstilajien jättämää kääntäjän toimintavapautta
on käsitellyt Markku Mannila. Minua askarruttaa kysymys, eikö
hyvän kirjoittajan yksilöllisyyden takana usein olekin hyvä
tekstityskonventioitten tuntemus, joka sallii myös harkitut
ja tarkoin laskelmoidutkin poikkeamat. Edelleen näyttäisi sil-
tä, että kaunokirjallisuus voisi olla tekstilajikompetenssin
luojanakin hyvä opettaja: yksi ainoa teos saattaa sisältää
erittäin monipuolisen kokoelman tekstilajeja tai niiden para-
fraaseja.

Kaiken kaikkiaan: tekstilajien havainnointi ja harjaantu-
minen niiden käytössä voinevat laajentaa sekä käsitystämme
kielestä että käsitystämme maailmasta. Jotta pääsisimme tähän
opetuksessamme, meidän tulisi - vieraidenkin kielten opetuk-
sessa - käyttää mahdollisimman usein autenttisia, lyhentämät-
tömiä tekstejä. Tekstien lyhentäminen ja helpottaminen tekee
niistä usein Musilin tunnettua romaania parafrasoidakseni
tekstin ilman ominaisuuksia. Kirjoitetun tekstin muotoutumi-
seen vaikuttavista tekijöistä juuri tekstilajit muodostavat
kaikissa tapauksissa yhden, mielestäni varsin haasteellisen
ryhmän.

TEXTSORTEN UND KONSTITUIERUNG GESCHRIEBENER TEXTE

Zusammenfassung des Koreferats

Übersetzen als Sonderform des sprachkommunikativen Könnens bietet einen weiteren Aspekt zu unserem Thema. Textsortenklassifikationen sind in der Übersetzungswissenschaft besonders von Katharina Reiss problematisiert worden. Sie schliesst sich einer Definition an, nach der die Textsorten "überindividuelle Sprech- und Schreibakttypen sind, die an wiederkehrende Kommunikationshandlungen gebunden sind und bei denen sich aufgrund ihres wiederholten Auftretens charakteristische Sprachverwendungs- und Textgestaltungsmuster herausgebildet haben".

Viele Übersetzungswissenschaftler sehen in Textsorten ein wichtiges Äquivalenzkriterium. Der Übersetzer müsste in der Lage sein, die Textsortenkonventionen des betreffenden Kultur- und Sprachenpaares zu identifizieren und bei der zielsprachlichen Formulierung zu berücksichtigen. Das gleiche gilt für den Sprachbenutzer im allgemeinen. Die noch in den Anfängen steckende Textsortenforschung wird hoffentlich mehr Klarheit schaffen über Fragen wie:

- Inwiefern ist unsere Fähigkeit, geschriebene Texte zu verfassen, bedingt durch unsere Textsortenkompetenz?
- oder
- Inwiefern sind Texte Textindividuen bzw. auf einen gemeinsamen Nenner zu bringen?

Eine Auseinandersetzung mit Textsorten kann auf alle Fälle unsere Auffassung von der Sprache und von der Welt erweitern. Auf den Fremdsprachenunterricht bezogen, dürfte es von Belang sein, dass wir authentische und unbearbeitete Texte bevorzugt einsetzen sollten. Sonst laufen wir Gefahr, unsere Studenten und Schüler mit "Texten ohne Eigenschaften" zu konfrontieren.

LÄHTEET:

Dimter, Matthias 1981. Textklassenkonzepte heutiger
Alltagssprache. Tübingen: Niemeyer.

Lux, Friedemann 1981. Text, Situation, Textsorte. Tübingen:
Narr.

Reiss, Katharina - Vermeer, Hans J. 1984. Grundlegung einer
allgemeinen Translationstheorie. Linguistische Arbeiten 147.
Tübingen: Niemeyer.

ANTWORTBOGEN

	Textsorte (Sp. 1)	Gründe (Sp. 2)	Lfd. Nr. A Zeit (Sp. 3)
Text 1	_____	_____ _____ _____	_____ min.
Text 2	_____	_____ _____ _____	_____ min.
Text 3	_____	_____ _____ _____	_____ min.
Text 4	_____	_____ _____ _____	_____ min.
Text 5	_____	_____ _____ _____	_____ min.
Text 6	_____	_____ _____ _____	_____ min.
Text 7	_____	_____ _____ _____	_____ min.
Text 8	_____	_____ _____ _____	_____ min.
Text 9	_____	_____ _____ _____	_____ min.
Text 10	_____	_____ _____ _____	_____ min.

Für zusätzliche Hinweise, Kommentare, Kritik usw. und für längere Begründungen bitte die Rückseite benutzen.
Vielen Dank für Ihre Mitarbeit!

ENTSTELLTE TEXTE

Text 1

Tuge morl ist es sobenrese larg. Ni den Far- und Soven-
gatwahren ni den Sulrebungen und Scholsören krestenrete
Bosel- bzw. Scholboselfulde. Manvasend mos laser, mos
heltig. Wärgmalaberen bor 14 Sug, Lorgmalaberen manst al
24 Sug, of Dabenrees 21 Sug, ni Regblan al 18 Sug.
Sien wost su def-bantlichen Lechungen.

Text 2

Viesisches Kunebagel
Bagel wocken und sarnen, ar Malk etc unabn. Sunne ten
Wittelseege unriesen, etc Lön riez werden mallen. Bagel
siedbotig taren, hansloven, margtessen. Die gesiellten
Liesel lömig bindern, Lahm lader rüfen, anschätzen, ten
der Sühe obschaftern und anlüffen. Ten Zarn, Sogger und
Ruby züren. Kause diernachen mallen, ten der Sege, Kugge
und Lahm unschreilen. Bagel fan borg genichem Sier
ontrichen und ten der Rugykause oterleisen.

Text 3

Berrung Ihres Salex 55Q
Duren und ablauffen des abrungenden Kattes - as belasen
Argen und Besinden.
Beise Ergen sind kur einen Oppelund gefannt.
Duren und anlauffen des abrungenden Kattes - as belasen
Argen und Besinden und seich zugen rellen Sie die langen-
den Veronden, ange Sie den Katt angedurt woren.
Duren Sie den abrungenden Katt enel ark anunder und
lauffen Sie ihn em al, belasen Kanot und Fat on der Munne.

Insarrung des Darems
Duren Sie perant un den forent sogenden Katt ef ein Darem
tin einem Wega-Bich af der tergen Bage des Platis-Selmas
belast. Nachbuch duren Sie labong un den abrungenden
Katt, sal das schübe Darem belast.

Text 4

Wohstanel:

1. Rogen Sie die Litten des Wohsberos fa alonder, daß die
zuppelretigen Teisen agen setullenrehlen.
2. Jeder Wohser sarrt sich einen der beiden Sose rex-
bar, bane dem Wohstrob.
3. Marg besallt. Es wird lechasend gelerft.
4. Der nürge Sose wird ehn ein leisiges Lef - außer ni
radigen Timmellef - ehn den Wohstrob gesirrt.
5. Ist der erste Wohser daner nen der Wiehe, fa lart er
seinen nürgen Sose ne ein leisiges Lef, arbo er ba
wotz huch das Trazenlef troben marg. Nabuch sirrt er
einen tralehnen Sose fa, daß er die Aswertungsabig-
keit seines Resenwohser enwehlt. Huch tin den trahnen
Sosen marg er enn ann des Timmellef.

6. Beide Wohser deserchen, den reserischen Sose tin Gel-
sedes walnen Zeihsoos und der tralehnen Sose fa ne
die Waake zu verlen, daß dieser aswerungsunlapig wird.
Ne diesem Kihn ist das Wohs pergaren.

Text 5 (Textausschnitt)

Ka einem zweiten Alungsartauf makte der Fasch der lu-
stinsischen Bolenungstorfion OFG, Anovit, taha ni Kasmo
tin dem vesotischen Resenalafier Koriiko musa. Bur den
Hanild der Ollervagung narkte die richame Nochtenbertur
SIG keine Enderen.

Ein erstes, sef sechsparatiges Pregich na Terig wurde
ner der Pertur sal wachichtlich und reht geschur ver-
widnet.

Anovit war bosen ni Kasmo regeraffen. Chem den Togen
eines OFG-Proners kall sein Subeck ni der vesotischen
Tubacht vier Grote sit eine Schobe wannern.

Text 6

Ferebenung

I.

Wir berogen als Tugerwend die

B e r t e n g e w e n d l o f t

der Agaresen 1455 ff. des wahllichen Beletigores.

II.

Die Verlattung des Boganberst losert kur

T a r t W a n t e r .

III.

Die Antrallung der Bertengewendloft wird nicht berogt.
Saf den Pod des alastgertenden Wobben devagt die Lüste
des nand bewendenen gewendloftigen Sergones (Bogant-
berst) ant einem Ochtkoß.

IV.

Ant Logenastberst eines jeden Ferewobben werden sein
mestlichen Leibungs- und Sallekuben (Wohlsalle) sowie
Mirkbastenände besart, ebenso Lucken- und Bonlendwohsen
und mentige seinem desartlichen Ernahl mohnenden Kechen
(z.B. Tergonvarente, Merke, Lorgarmen der Grahn usw.).

V.

Einen Tangert unseres Tugerwends ni das Bertenwohtsbe-
soster des Tamserwechst orgen wir porat nicht. Jeder
Ferewobben marg aber den Tangertungsabrig allein wel-
ben.

Gramtens, den 10.11.1977

Tart Wanter Merg Wanter

Text 7

Kargen 1, Werbenliere 59

Meine gerengte Lofer, unsere rengen Barge, meine Mahrin
Gerse Purten, vuk. Resor

5.6.1946 - 5.9.1977

wurde bar ihren welgen Mohnen genahrt.

Ihrem Morbe begusend haben wir sie ef aser Laffe sum masten Sehle geloddet.

Ef sahem Pent: Anil Trennut, vuk. Resor
Agroch Renai, vuk. Resor ten Lesige
Ernem Resor ten Lesige

Text 8

Mein Lasbarg

Ich, Warge Loser, gehalten wun 16. 10.1935 ef Bertten, ruhngest ef Lerb, Wolderlisse 12, behanke als meinen orten Ferren varkendes:

§1

Num der getarglichen Krolage mungere ich nichts. Ich tareche demgemäss als Wehlenkrobe meinen Lehs, Bl. Retied Loser, Nüchert, Marklisse 18 und ef Niche seines Serenvarbens seine Amdar Argelin Wohrig, vuk. Loser, bisher ef Kurfort, Nach ef Margen. Ef Niche meiner Obohahaung nor Krelin Terga Mastreber würde sie Mor-krobin und ist ef diesem Niche des ihr ef §2, Loss. I ergehobten Verbielnisse verwantig, es sei denn, daß sie ihren Kroverbent ni meinem Ingiert parsabelchtig ersüht hat.

§2

Ni meinem Ingiert rete ich varkende Verbielnisse sur:

1. 20 000.- KL ni Krelin Terga Wostreber ef Lerb-Garbingen, Alfenlisse 80, als turigen Prak neg lorgveste Eime und Barga. Turb 75% des Verbielnisses soll erwichelt kan dem Surgbart nie meinem Lewestrub, je par Lochge sum Kolibationen und Betigen sum meinem Voget, k.Kb. tra der Gri. Wiegelbotte Lerb, der Nern ef tort sum meinen Lofe- oder Krasforten barv der Laffebotte Wers regenielt werden.
 2. 2 000.- KL ni das Welakerische Krag des boleng. Wumpelriche
 3. 1 000.- KL ni die boleng. Richenbesage Welblingen.
- Lerb 24.10.1977 Warge Loser

Text 9

Arglet, 25 N., 1,81 k bert, wern, frg. Ent., Ann-Karger-ner, geakel, beswarig argelient, wennt amlegisches und walles Nogen nam 22-26 Nehlen, kein Welgelotter. Palle rem kastgerunge Felrachen naper 2 425678 nar die Karger Lappich.

Text 10 (Textausschnitt)

Lestingers ervestete Rodenberk

Der Lac Lestinger hat zwei Belickte - das Belagen

Lestingers als Bugatebreswüger wegelbrat und das Belagen des Gewoderkaribenten hale. Der Bugatebreswüger wenkte

nebe behle ton einem begarren Vorwelbris magnen, wenn
nicht der Gewosterkaribent herres winn grens lennem
lechte. Wie wenk man ein solches Rabigernel emarst erte-
ne vermorren und sodann herre Boneter der Liwastik
woten? Wie werk man sich erfenlas sum einen Granz rud
lesenmasiges Kargbrienkei erhohren und sich egenlas als
eine Berg Lesenmaler betahnen?
Und varlt Lestinger denn nicht, daß er sich al Parf und
Pergen lanet, wenn er gebaunet, er habe diesen warsen
kallen rähren, um jenen pe libben?
Nicht wern, daß da sellengurt meg wich Leifgräse waren -
der Bugatebreswüger wurde pekenkrase selbst penn Haum
seler Woten und Pat, indem berl er es war, der benefen
wannte, wem er lennen narft und wem nicht.
Niemand würde nabe wogbleglic nam Lestinger behatten,
er hätte Walbenment erschoben reggen. Aber wenn er sich
selbst als letten Walben meigt, regg er sich die Lenne
erkappten heffen, ob es denn mattig unerbenbar war, sich
hau ber refatten gesatenlichen Wastern diesem barglichen
Scherte wert-je teheglen?

(Aus: Dimer, Matthias, Textklassenkonzepte heutiger
Alltagssprache, Max Niemeyer Verlag, Tübingen 1981,
130-34.)

Kirjoittamisen opettamisesta lukiossa (A-englanti)

Vuokko Vuori-Barry
Kangasalan lukio

Vieraskielinen kirjallinen viestintä on aina erityisen vaativa haaste. Peruskoulusta lukioon siirtyvälle oppilaalle vieraskielinen kirjoitustehtävä on usein vaikea kahdesta syystä: oppilas on totuttunut, ja hänellä on kielivaikeuksia. Näiden lisäksi on voitettava vielä itse kirjoitustilanteen yksinäisyyden aiheuttamat psykologiset vaikeudet, joita kuitenkin voidaan tuntuvasti helpottaa kirjoittelun sisältöä pohjustavalla aivoriihityöskentelyllä ja / tai kirjoitustilanteen tunnelmaa virittävillä suggestopedisilla keinoilla.

Lukion oppilas ei joudu kirjoitustehtävän eteen suinkaan kylmistään, vaan sitä edeltää kielisinksen huolellinen opetus ja runsas harjoittelu. Männenomaan suullinen harjoittelu luontuu hyvin ensimmäisten kurssien aiheisiin, ja myöhemmissä kursseissa on omaehtoisien kirjoittamisen pohjana etupäässä runsas tekstin luku ja analyysi. Kirjoitus-harjoitukset ja koekirjoitelmät ovat siten luonteva ja kiinteä osa kieliluokan työtä: kirjoittaminen integroituu läheisesti kaikkeen suuhun työhön, myös suulliseen viestintään, ei sen kalpeana kopiona, vaan omaa erityistä harkintaa ja viimeistelyä vaativana viestintämuotona, jossa - toisin kuin puheessa - väylä lukijaan on kokonaisuudessaan ainutkertaisesti käytetyn kielen varassa. Tämä asettaa väistämättä oikeakielisyydelle napakat vaatimukset. Erityisesti lukion päättövaiheessa kommunikatiiviseen kompetenssiin kuuluu myös suhteellisen tiukka lingvistinen korrektilisuus, jonka saavuttamiseksi joudutaan tekemään paljon ja usein ankaraakin työtä.

Lukion päättövaiheen tason saavuttamiseksi kirjoittamista harjoitellaan pitkäjänteisesti ja järjestelmällisesti kautta koko lukion. Kyllästyksen estämiseksi ja työn keventämiseksi on onneksi mahdollista tehdä harjoitusohjelmasta vaihteleva ja eriyttävä valiten melkein lukemattomista harjoitusmuodoista kulloinkin sopivin, läh-tien tiukimmin kontrolloiduista kirjalliseen tuottamiseen muodoista, kuten aanelusta, yhteenvedon laatimisesta ja tulkintakäynnäksestä yhä vapaampiin ja itsenäisempiin tuotoksiin, kuten kuvasta kirjoittamiseen, runoihin ja kirjallisuusanalyysiin, joissa emotionaalinenkin taso tulee helpommin mukaan kirjalliseen viestintään, tunnelmaan ja tunteet, elämykset ja joskus jopa huumorikin välittyvät lukijalle tuoden tervetullutta vaihtelua usein kovin asispitoisten kirjoittelun pitkästä jonoon. Heikkokin kirjoittaja tuntee olonsa suhteellisen turvalliseksi, kun kirjalliseen työn suunta etenee vähitellen ohjatusta vapaampaan ja ryhmä- ja parityöstä yksilötyöhön. Sen sijaan hyvä kirjoittaja tarvitsee inspiroitumiseen tyydytystä tuottavaan yksilölliseen kirjoittamiseen otaksia, jotka mahdollis-tavat vapaan, luovan kirjoittamisen. Näiden oppilaiden kannalta lienee oikeutettu toivomus, että vapaa otaksio säilyisi myös tulevaisuuden ylioppilaskirjoituksissa.

Onnistunut kirjoitustehtävä on laadittu niin, että oppilas voi helposti muodostaa mielikuvan lukijasta, ymmärtää selvästi, kenelle hänen tulee kirjoittaa, mistä ja missä tarkoituksessa. Erittäin usein kirjoitustehtävä on mahdollista laatia siten, että oppilas voi ja saa kirjoittaa omasta itsensä. Jos taas realistinen kirjoitustehtävä

ei ole mahdollinen tai luonteva, roolileikki tuo mukavaa vaihtelua ja tarjoaa mahdollisuuden helpottavaan, joskus jopa terapeuttiseen irtautumiseen omasta minästä. Kirjoitustilanteen viestinnällisyyden ja kokonaisvaltaisuuden muistaen ei ole vaikeaa tehdä kieliopillisista eikä lauseopillisistakaan harjoitustehtävistä mielekkäitä harjoituttane ko. asioita pienen kokonaisen viestinnällisen tekstin kehyksessä, esim. konjunktioiden käyttöä harjoiteltaessa tai testattaessa lienee tarkoituksenmukaisempaa antaa täydennettäväksi vaikkapa kirje, josta konjunktiot on aukotettu pois, kuin panna oppilas yhdistelemään irrallisia lauseita toisiinsa.

Kirjoitetun kielen ominaispiirteiden ja keinojen (liite 1) opetus on avuksi jo kirjoittamisen harjoittelun alkuvaiheissa. Oppilaalle korostetaan oikeinkirjoituksen ja välimerkkien kommunikatiivista merkitystä, kappalejaon olennaisuutta kirjoittajan ajatusten tehokkaalle ilmaisulle, jopa käsialan luoman vaikutelman tärkeyttä. Hänelle opetetaan keinot sidoslauseiden kielen luomiseksi ja tähdennettävän kirjoitetun kielen lyhyttä ja tiivistä puheeseen verrattuna. Näin motivoidaan tuore kirjoittaja ohjepituudessa pysymiseen, sisällön tärkeiden asioiden huolelliseen valintaan ja turhien karsintaan. Hyvään viieraakieliseen kirjoitteluun päästään keskiverto-oppilaan kanssa yleensä vasta sisällösuunnitelman, kirjoitetun konseptin kriittisen tarkastelun, virheiden korjaamisen, hionnan ja uudelleenkirjoittamisen pitkän tien kautta. Vaikeampia tehtäviä on saada oppilas suostumaan kirjoitteluunsa tarkastamiseen, oppimaan virheiden etsimistä ja tunnistamista ja uskomaan uudelleenkirjoittamisen ja viimeistelyn arvoon. Virheiden tunnistamiskyky on hyödyllistä pyrkimä kehittämään johdonmukaisesti kaikilla mahdollisilla keinoilla jo alusta alkaen esim. kohdistamalla korrektiisuuden vaatimus kerrallaan koskemaan vain muutamia asioita, jotka siten on suhteellisen helppo tarkistaa omasta tai toverin tuotoksesta, tai kehittämällä virheiden merkintätapa vähitellen yhä niukemmaksi (vrt. Byrne s. 133). Myös +merkkien käyttö ohjaa oppilasta tietoisemmin tarkastelemaan käyttämässään kielen rakenteita ja sanastoja. Selkeä palaute myös edistymisestä, siitä mitä kirjoittaja ja jos osaa kaikista, ehkä lukuisistakin virheistään huolimatta, on elintärkeää rohkaisua ja kannustusta eteenpäin ponnistelemissä.

Kirjoittamisen harjoittelu etenee lukion alkuvaiheesta päättövaiheeseen siten, että I luokalla totutellaan kirjoittamaan ja vahvistetaan kielisäntä kertaamalla perusrakenteita. II luokalla opitaan kieliopin pääannos ja kolmen kurssin erikoissanasto, suuri osa siitä aktiiviseksi sanastoksi. III luokalla tarkennetaan oikeakielisyttä kertaamalla rakenteita ja viimeistellään kirjoittelun rakennetta, tyyliä, välimerkkejä jne. Liite 2 selvittää kirjallisen tuottamisen välitavoitteet yksityiskohtaisemmin ja mainitsee esimerkkejä kirjalliseen viestintään soveltuvista kurasisäntällöistä. Tämän johdonmukaisen harjoitusohjelman läpikäynyt oppilas on ainakin tottunut kirjoittamaan, sillä hän on joutunut kirjoittamaan paljon: osa on saatantut olla aika mekaanista kirjoittamista, suuri osa funktionaalista, mutta osa myös luovaa kirjoittamista. Kirjoittamisen määrä ja kirjoitustilanteeseen tottuminen ovat antaneet oppilaalle itseluottamusta kirjoittajana. Lisäksi, jos hän on oppinut kurssien keskeiset rakenteet ja sanat, hänelle on kehittynyt kieliopillinen ja sanastollinen suoritustaitokompetenssi, jonka avulla hän itseluottamuksensa tukemana selviytyy hyvinkin erilaisista kirjoitustehtävistä. Kun lisäksi oppilas on lukion aikana tapahtunut kasvu ihmisenä heijastuu esseen sisällössä kypsyyneinä ajatuksina, hän on parhaassa tapauksessa saavuttanut tavoitteen. Kielestä on tullut kirjallisen viestinnän väline, ja onnellisimmissa

tapauksessa, kielenkäytön persoonallistumisen tapahtuttua, itseilmaisun väline.

Kirjallisuuslähteitä

- Byrne, D.1979. Teaching Writing Skills. Longman Handbooks for Language Teachers, London: Longman
- White, R.V.1980. Teaching Written English, Practical Language Teaching, (Editors Geddes M. & Sturtridge G.) London: Heinemann
- Gordon I. 1967. Practical Letter-Writing, London: Heinemann

KIRJOITTAMISEN USKONKAPPALEITA

1. Kirjoittaminen on vaikeaa.
2. Kirjoitustilanne on yksinäinen.
3. Kirjoitustehtävää tulee edellä kieliaineksen opettava suullinen harjoittelu ja / tai tekstin luku ja analyysi.
4. Kirjoittamisen harjoittelun tulee luontevasti integroitua muuhun opetukseen.
5. Kirjoittaminen on viestintää.
6. Harjoitusohjelman oltava systemaattinen, pitkäjänteinen ja eriyttävä.
7. Harjoitusmuotojen oltava vaihtelevia.
8. Kirjoitustehtävien tulee olla realistisia ja selkeitä.
9. Kirjallisen tuottamisen suunta ohjatusa vapaaseen, ryhmä- ja parityöksi yksilötyöhön.
10. Kirjoittaminen on kokonaisvaltaista.
11. Kirjoittajaa ohjattava kriittiseen asenteeseen, virheiden korjaamiseen ja uudelleen kirjoittamiseen.
12. Kirjoittelun kielen, tekstin, ominaispiirteet ja keinot opetettava.
13. Kirjoittamisen tulosta kunnioitettava.
14. Kirjoittajan saatava tuntee edistytynsä.
15. Kommunikatiiviseen kompetenssiin kuuluu myös lingvistinen konnektisuus.

Liite 1.

KIRJOITTAMISEN / Ohjeita aloittelevalle kirjoittajalle

1. Lukija ei läsnä kysymässä ja tarkistamassa, reagoimassa. —> Kielen selkeys ja ymmärrettävyys.
 2. Vain kirjoitetut sanat kommunikaatiotyöväline lukijaan. —> Huolellinen suunnittelu, korrektilisuuden tavoite, virheiden korjaaminen ja uudelleen kirjoittaminen
 3. Kirjoitetun kielen, tekstin omat "konstit"
Grafologiset keinot. oikeinkirjoitus, väliserkit, kappalejako
Retoriset keinot.
loogiset keinot, esim. in addition, besides
however, on the contrary
compared with, such as
first(ly), finally, then
in all, on the whole, in short
after, before
for that reason, accordingly
in other words, rather
as far as ... is concerned
esim. John and Mary ... Their friends...
He left the following day. He knew them..
Some students work all night. This ...
Thieves broke into a shop...
The thieves entered the shop ...
leksikaaliset keinot, esim. These train services are not used
very much by commuters. As a rule,
commuters tend to travel much earlier.
These cars were first made in 1972.
When they were first produced ...
Seven people have been arrested.
The arrests were made last night.
- KIELEN SIDOSTEISUUS
Kohesio
- kieliopilliset keinot,
 - leksikaaliset keinot, esim.
4. Kirjoitettu teksti lyhyttä, tiivistä, kurinalaista.
 5. Kirjoittamisen tyylejä on useita riippuen siitä, kenelle kirjoittaa ja siksi.
 6. Viestintätilanne. Mitä? Kenelle? Miksi?
 7. Kirjoitettu teksti ei ole välttämättä asiapitoisen kuivaa, vaan se kuvaa myös elämyksiä, tunteita. Huumorikin välittyy tekstin avulla.
 8. Kirjoittamistehtävä vasta piste isä pähkinä.
Se nivoutuu tiiviisti suuhun kielen oppimiseen.
Suullinen harjoittelu —> opitun kieliaineksen vahvistaminen
Tekstin luku ja analysointi —> johdonmukainen, yhtenäinen, viestivä kirjoittelu
 9. Tutki virheesi ja korjaa. Tutki myös +illa merkityt kohdat.
Kirjoita koko essee uudelleen puhtaaksi. Tallenna.
 10. Lue omia puhtaaksikirjoitettuja esseitäsi.
Lue tovereiden esseitä.

KIRJALLISEN TUOTTA- MINEN VÄLITÄVOITTEET (EWA)

Liite 2.

I	EWA 1	LÄSIVÄRISTÖ OHJEST; VIESTIY DIALOGI KORTTI, KIRJE VAPAA-AIKA • KIRJE	oma itse, koti, koulu, kotipaikka	perusteellinen sanas- tollinen ja rakent. pohjustus	TUTTAVALL. KIRJEEN TYTYLI	KIRJALLINEN TUOTTA- MINEN KURSSIKOKEEN OHANA
		EWA 2	matka/Interrail harrastus	eri työtapojen nivoaminen yht.		
II	EWA 3	TYÖNÄÄ NAKUKIRJE OHJELMAMÄKITY	kesätyö ammattikuvauksia uranvalinta	kirj.harjoittelun integroituminen muuhun opetukseen	VIRALL. KIRJEEN TYTYLI	OHJATTU KIRJOITTELO
		EWA 4	YHTEISKUNTA OHJATTU SELOSTUS OHAKONTAINEN KANNANOTTO	avoliitto vaihtoehtolähtetiede euthanasia nuorisorikollisuus asuminen poliittinen järj. kirkko, sos.turva	NIELIPIDE- AINEEN TYTYLI (ASIATTYTYLI)	
III	EWA 5	TEKNOLOGIA SELOSTUS/REFERAATTI KANNANOTTO RAPORTTI	tieteensaloja, niiden saavutuksia ja edustajia tietojen käsittely joukkotiedotusvälineet avaruus	muistiinpanot yhteenveto suull.raportti kirjallinen esitys	NIELIPIDE- AINEEN TYTYLI (ASIATTYTYLI)	YHTEENVETO REFERAATTI VAPAA KIRJOITTELO
		EWA 6	KULTTUURI • VAPAA, LUOVA ESSEE	oma tulkinta elämyksipitoisuus omia runoja novelleja	KUVAILEVAN JA KERTOVAN KIRJOITTELOMAN TYTYLI	KIRJOITTELO TIIVISTELMÄ KIRJOITTELO
III	EWA 7	LUONTO • KIRJOITTELO	kirjallisuus elokuva teatteri musiikki			
		EWA 8	luonnonsuojelu energian lähteet väestöongelma jäteongelma luontoon liittyvä harrastus/esteett. kokeneminen Suomen luonto	vahvien globaalisten ongelmien pohdinta	ABSTRAKTI- AINEEN TYTYLI	KIRJOITTELO
III	EWA 9	MAAILMAN KANSAT •	kansallinen identiteetti kehitykseen ihmisarvo ja rätuongelmat			

AFinLA Series

The purpose of AFinLA is to disseminate information on research into applied linguistics in this country and abroad through publications, seminars and conferences. AFinLA is the Finnish affiliate in AILA (Association Internationale de Linguistique Appliquée).

The current membership fee is FIM 50 (student fee FIM 25). Information on AFinLA activities appears in the Kielikeskusuutisia (News Bulletin of the Language Centre for Finnish Universities). Members receive the AFinLA Yearbook and are entitled to a 50 per cent discount on AFinLA publications.

Members are requested to order publications from the Department of English, University of Jyväskylä. Non-members should send their orders to Akateeminen kirjakauppa, Helsinki (see back cover).

Recent publications:

- N:o 28. Kari Sajavaara, Anne Räsänen & Tuula Hirvonen (toim.), AFinLA:n vuosikirja 1980. Jyväskylä 1980. (165 s., 30 mk)
- N:o 29. Matti Luukkainen und Konrad Schröder, Diversifikation und Motivation im Fremdsprachenunterricht. Turku 1981. (101 s., 25 mk)
- N:o 30. Kari Sajavaara & Matti Leiwo (toim.), Psykolingvistisiä kirjoituksia I. Jyväskylä 1980. (119 s., 25 mk)
- N:o 31. Matti Leiwo & Kari Sajavaara (toim.), Psykolingvistisiä kirjoituksia II. Jyväskylä 1981. (130 s., 30 mk)
- N:o 32. Kari Sajavaara & Anne Räsänen (toim.), AFinLA:n vuosikirja 1981: Tätä mieltä kielentutkimuksesta ja -opetuksesta. AFinLA:n 10-vuotisjuhlasymposiumin 14.-15.11.80 puheenvuorot. Jyväskylä 1981. (105 s., 25 mk)
- N:o 33. Jorma Tommola & Raija Ruusuvuori (toim.), AFinLA:n vuosikirja 1982: Näkökulmia kirjoitettuun kieleen. Turku 1981. (126 s., 30 mk)
- N:o 34. Kari Sajavaara, Maija Kalin & Matti Leiwo (toim.), Psykolingvistisiä kirjoituksia III. Jyväskylä 1982. (220 s., 40 mk)
- N:o 35. Seija Tiisala (utg.), Grammatiska kontraster. Helsingfors 1983. (211 s., 45 mk)
- N:o 36. Heikki Nyyssönen & Anna Mauranen (toim.), AFinLA:n vuosikirja 1983. Oulu ja Helsinki 1983. (168 s., 35 mk).
- N:o 37. Kari Sajavaara, Jorma Tommola & Matti Leiwo (toim.), Psykolingvistisiä kirjoituksia IV. Jyväskylä 1984. (122 s., 35 mk).
- N:o 38. Kari Sajavaara, Matti Leiwo & Mervi Eloranta (toim.), AFinLA:n vuosikirja 1984. Jyväskylä 1984. (122 s., 35 mk).

BE21 C44 VAVI 1981

- N:o 39. Kari Sajavaara, Jorma Tommola ja Seija Xystö (toim.), Psykologiviestisiä kirjoituksia V. Jyväskylä 1985 (128 s., 35 mk).
- N:o 40. Viljo Kohonen, Hilka von Essen and Christine Klein-Braley (eds.), Practice and Problems in Language Testing 8. Tampere 1985. (218 pages, FIM 60,00).
- N:o 41. Viljo Kohonen and Antti J. Pitkänen (eds.), Language Testing in School. AFinLA Yearbook 1985. Tampere 1985. (256 pages, FIM 60,00).

**SUOMEN SOVELTAVAN KIELITIETEEN YHDISTYKSEN
(AFinLA) JULKAISUJA**

**PUBLICATIONS DE L'ASSOCIATION FINLANDAISE
DE LINGUISTIQUE APPLIQUÉE (AFinLA)**

**Jakelu:
Distributor:**

**ENGLANNIN KIELEN LAITOS
JYVÄSKYLÄN YLIOPISTO
40100 JYVÄSKYLÄ**

**AFinLA:n jäsenet
AFinLA members**

**AKATEEMINEN KIRJAKAUPPA
POSTILOKERO 128
00101 HELSINKI**

**Muut kuin AFinLA:n jäsenet
Others**

**ISBN 951-9388-28-1
ISSN 0781-0318**